

28th European Conference on Modelling and Simulation

May, 27th - 30th, 2014, Brescia, Italy

ECMS 2014

Edited by:

Flaminio Squazzoni

Fabio Baronio

Claudia Archetti

Marco Castellani

Honorary Chairs:

Andrzej Bargiela

Eugène J.H. Kerckhoffs



Copyright

© ECMS2014

Printed: ISBN: 978-0-9564944-8-1

**European Council for Modelling
and Simulation**

CD: ISBN: 978-0-9564944-9-8

Cover picture

© ECMS

Printed by

**Digitaldruck Pirrot GmbH
66125 Sbr.-Dudweiler, Germany**

PROCEEDINGS

28th European Conference on Modelling and Simulation ECMS 2014

May 27th – May 30th, 2014
Brescia, Italy

Edited by:

Flaminio Squazzoni

Fabio Baronio

Claudia Archetti

Marco Castellani

Honorary Chairs:

Andrzej Bargiela

Eugène J.H. Kerckhoffs

Organized by:

ECMS - European Council for Modelling and Simulation

Hosted by:

University of Brescia, Italy

Sponsored by:

University of Brescia, Italy

Linea Com, Italy

COMSOL, Italy

International Co-Societies:

IEEE - Institute of Electrical and Electronics Engineers

ASIM - German Speaking Simulation Society

EUROSIM - Federation of European Simulation Societies

PTSK - Polish Society of Computer Simulation

LSS - Latvian Simulation Society

ECMS 2014 ORGANIZATION

Conference Chair

Flaminio Squazzoni

University of Brescia
Italy

Conference Co-Chair

Fabio Baronio

University of Brescia
Italy

Programme Chair

Claudia Archetti

University of Brescia
Italy

Programme Co-Chair

Marco Castellani

University of Brescia
Italy

President of European Council for Modelling and Simulation

Evtim Peytchev

Nottingham Trent University
United Kingdom

Managing Editor

Martina-Maria Seidel

St. Ingbert
Germany

INTERNATIONAL PROGRAMME COMMITTEE

Agent-Based Simulation

Track Chair: **Michael Möhring**
University of Koblenz-Landau, Germany

Co-Chair: **Ulf Lotzmann**
University of Koblenz-Landau, Germany

Simulation in Industry, Business and Services

Track Chair: **Alessandra Orsoni**
University of Kingston, United Kingdom

Co-Chair: **Serhiy Kovala**
University of Kingston, United Kingdom

Co-Chair: **Arne Petermann**
Free University of Berlin, INA, Germany

Simulation of Intelligent Systems

Track Chair: **Zuzana Kominková Oplatková**
Tomas Bata University of Zlín, Czech Republic

Co-Chair: **Roman Senkerik**
Tomas Bata University of Zlín, Czech Republic

Finance, Economics and Social Science

Track Chair: **Barbara Dömötör**
Corvinus University of Budapest, Hungary

Co-Chair: **Flaminio Squazzoni**
University of Brescia, Italy

Simulation of Complex Systems & Methodologies

Track Chair: **Krzysztof Amborski**
Warsaw University of Technology, Poland

Co-Chair: **Jaroslav Sklenar**
University of Malta, Malta

Simulation, Experimental Science and Engineering in Maritime Operations

Track Chair: **Hans Petter Hildre**
Aalesund University College, Norway

Co-Chair: **Sashidharan Komandur**
Aalesund University College, Norway

Electrical and Electromechanical Engineering

Track Chair: **Sergiu Ivanov**
University of Craiova, Romania

Co-Chair: **Vittorio Ferrari**
University of Brescia, Italy

Co-Chair: **Maria José Resende**
TU Lisbon, Portugal

Simulation and Visualization for Training and Education

Track Chair: **Vilmar Æsøy**
Aalesund University College, Norway

Co-Chair: **Houxiang Zhang**
Aalesund University College, Norway

Modelling, Simulation and Control of Technological Processes

Track Chair: **Jiří Vojtěšek**
Tomas Bata University in Zlín, Czech Republic

Co-Chair: **Petr Dostál**
Tomas Bata University in Zlín, Czech Republic

Co-Chair: **František Gazdoš**
Tomas Bata University in Zlín, Czech Republic

Discrete Event Modelling and Simulation in Logistics, Transport and Supply Chain Management

Track Chair: **Gaby Neumann**

Technical University of Applied Sciences Wildau, Germany

Co-Chair: **Edward J. Williams**

University of Michigan-Dearborn, USA

Policy Modelling

Track Chair: **Emile Chappin**

Wuppertal Institute for Climate, Environment and Energy,
Delft University of Technology, The Netherlands

Co-Chair: **Flaminio Squazzoni**

University of Brescia, Italy

Simulation and Optimization

Track Chair: **Frank Herrmann**

University of Applied Sciences Regensburg, Germany

Co-Chair: **Erik Krobot**

University of Bundeswehr in Munich, Germany

Co-Chair: **Thorsten Claus**

International Graduate School (IHI) Zittau, Germany

Co-Chair: **Claudia Archetti**

University of Brescia, Italy

Modeling and Simulation in Robotic Applications

Track Chair: **Sigal Berman**

Ben-Gurion University of the Negev, Israel

Co-Chair: **Sanja Dogramadzi**

Bristol Robotics Laboratory

University of the West of England, United Kingdom

High Performance Modelling and Simulation

Track Chair:

Joanna Kolodziej

Institute of Computer Science Cracow University of Technology, Poland

Co-Chair:

Horacio-Gonzalez-Velez

National College of Ireland Dublin, Ireland

Modelling and Simulation of Data Intensive Systems - Special Session

Track Chairs:

Joanna Kolodziej

Institute of Computer Science Cracow University of Technology, Poland

Ciprian Dobre

University Politehnica of Bucharest, Romania

Florin Pop

University Politehnica of Bucharest, Romania

Research and Use of Multiformalism Modeling Methods - Special Session

Track Chairs:

Marco Gribaudo

Politecnico di Milano, Italy

Mauro Iacono

Seconda Università degli Studi di Napoli, Italy

Probability and Statistical Methods for Modelling and Simulation of High Performance Information Systems - Special Session

Track Chairs:

Alexander I. Zeifman

Vologda State Pedagogical University, Russia

Pavel O. Abaev

Peoples' Friendship University of Russia, Russia

Rostislav V. Razumchik

Peoples' Friendship University of Russia, Russia

IPC Members in Alphabetical Order

Krzysztof Cetnarowicz, AGH University of Science and Technology, Poland

Frederic Amblard, University of Toulouse 1, France

Piotr Arabas, Warsaw University of Technology and NASK, Poland

Monika Bakosova, Slovak University of Technology in Bratislava, Slovakia

Hans-Peter Barbey, University of Applied Sciences in Bielefeld, Germany

Fabio Baronio, University of Brescia, Italy

Avital Bechar, Volcani Center, Israel

Nik Bessis, University of Derby, United Kingdom

Shusheng Bi, Beihang University, China

Armin Biess, General Motors-R&D, Israel

Froy Birte Bjorneseth, Rolls Royce Marine, Norway

Riccardo Boero, Los Alamos National Laboratory, USA

Fabian Boettinger, Fraunhofer Institute IPA, Germany

Giangiacomo Bravo, University of Linnaeus, Sweden

Øyvind Bunes, Rolls-Royce Marine, Norway

Aleksander Byrski, AGH University of Science and Technology, Poland

Piers Campbell, University of Newcastle NSW, Australia

Petr Chalupa, Tomas Bata University in Zlín, Czech Republic

Marielle Christiansen, NTNU Trondheim, Norway

Catherine Cleophas, RWTH Aachen, Germany

Daniele Codetta Raiteri, University of Piemonte Orientale, Italy

Valentin Cristea, University Politehnica of Bucharest, Romania

Peter Csóka, CUB, Hungary

Gregoire Danoy, University of Luxemburg, Luxemburg

Donald Davendra, VŠB-Technical University of Ostrava, Czech Republic

Roman Debski, AGH University of Science and Technology, Poland

Bruno Dehez, University Catholique of Louvain, Belgium

Bernabe Dorronsoro, University of Lille 1, France

František Dušek, University of Pardubice, Czech Republic

Andrzej Dzielinski, Warsaw University of Technology, Poland

Bruce Edmonds, Univ. Business School Manchester, United Kingdom

Daniel Fürstenau, Free University of Berlin, Germany

Simone Gabbriellini, University of Paris-Sorbonne, France

Charlotte Gerritsen, VU University of Amsterdam, The Netherlands
Dag. Sverre Gronmyr, Rolls Royce Marine, Norway
Alexander A. Grusho, Russian Academy of Science, Russia
Daniel Grzonka, Cracow University of Technology, Poland
Karl Henning Halse, Aalesund University College, Norway
Thomas Hellström, Umea University, Sweden
Gerald Holowicki, University of Michigan – Dearborn, USA
Daniel Honc, University of Pardubice, Czech Republic
Mark Hoogendorn, VU University of Amsterdam, The Netherlands
Thomas Hußlein, OptWare GmbH in Regensburg, Germany
Martin Ihrig, University of Pennsylvania, USA
Teruaki Ito, The University of Tokushima, Japan
Michal Janosek, University of Ostrava, Czech Republic
Cara Kahl, Technical University Hamburg-Harburg, Germany
Bogumil Kaminski, Warsaw School of Economics, Poland
Uri Kartoun, Harvard Medical School, USA
Eugène J.H. Kerckhoffs, Delft University of Technology, The Netherlands
Stefan Klaußner, Europe-University Viadrina Frankfurt (Oder), Germany
Natalia Kliewer, Free University of Berlin, Germany
Petia Koprinkova, Bulgarian Academy of Sciences, Bulgaria
Victor Yu. Korolev, Lomonosov Moscow State University, Russia
Igor Kotenko, SPIIRAS, Russia
Martin Kotyrba, University of Ostrava, Czech Republic
Imola Kovács, Babes-Bolyai University, Romania
Marek Kubalčík, Tomas Bata University in Zlín, Czech Republic
Jane Labadin, University Malaysia Sarawak, Malaysia
Kai Lampka, Uppsala University, Sweden
Alex Lenz, Bristol Robotics Laboratory, United Kingdom
Catalin Leordeanu, University Politehnica of Bucharest, Romania
Wei Li, Aalesund University College, Norway
Guoyuan Li, Aalesund University College, Norway
Dario G. Liebermann, Tel-Aviv University, Israel
Rong Liu, Beihang University, China
Ahmad Lotfi, Nottingham Trent University, United Kingdom

Dorin Lucache, "Gh.Asachi" Technical University of Lasi, Romania
Michael Manitz, University Duisburg-Essen, Germany
Michal Marks, Research and Academic Computer Network (NASK), Poland
Michael Mäs, ETH Zurich, Switzerland
Katarina Matějíčková, VUCHT, Slovakia
Galina Merkurjeva, Riga Technical University, Latvia
Nicolas Meseth, Deloitte Consulting, Germany
Ruth Meyer, Manchester Metropolitan University Business School, United Kingdom
Benjamin Michels, International Academy at Free University of Berlin, Germany
Dan Mihai, University of Craiova, Romania
Christian Müller, University of Applied Sciences in Wildau, Germany
Pavel Nahodil, Czech Technical University in Prague, Czech Republic
Catalin Negru, University Politehnica of Bucharest, Romania
Ewa Niewiadomska-Szynkiewicz, Warsaw Univ. of Technology and NASK, Poland
Libero Nigro, University of Calabria, Italy
Jakub Novak, Tomas Bata University in Zlín, Czech Republic
Ilie Nuca, Technical University of Moldova, Republic of Moldova
Dominik Olszewski, Warsaw University of Technology, Poland
Ottar Osen, Aalesund University College, Norway
Bernd Page, University of Hamburg, Germany
Teodor Pana, Technical University Cluj-Napoca, Romania
Mario Paolucci, ISTC/CNR, Italy
Alexander V. Pechinkin, Russian Academy of Science, Russia
Tony Pipe, Bristol Robotics Laboratory, United Kingdom
Jeremy Pitt, Imperial College, United Kingdom
Michal Pluhacek, Tomas Bata University in Zlín, Czech Republic
Gary Polhill, James Hutton Institute, United Kingdom
Ioan Popa, University of Craiova, Romania
Francesco Pupo, University of Calabria, Italy
Simone Righi, Hungarian Academy of Sciences, Hungary
Boris Rohal-Ilkiv, Slovak University of Technology in Bratislava, Slovakia
Wasko Rothmann, Europe-University Viadrina Frankfurt (Oder), Germany
Juliette Rouchier, GREQAM/CNRS, France
Konstantin E. Samouylov, Peoples' Friendship University of Russia, Russia

Roberto San José, Technical University of Madrid (UPM), Spain
Johnatan Pecero Sanchez, University of Luxemburg, Luxemburg
Sabrina Scherer, University of Koblenz-Landau, Germany
Sergey Ya. Shorgin, Russian Academy of Science, Russia
Nir Shvalb, Ariel University Center, Israel
Peer-Olaf Siebers, The University of Nottingham, United Kingdom
Alexander Simon, Berlin University for Professional Studies, Germany
Anders Skoogh, Chalmers University of Technology, Sweden
Andrzej Stefan Sluzek, Khalifa University Abu Dhabi, UAE
Katarzyna Smelcerz, Cracow University of Technology, Poland
M. Grazia Speranza, University of Brescia, Italy
Matthew Studley, Bristol Robotics Laboratory, United Kingdom
Lorand Szabo, Technical University Cluj-Napoca, Romania
Magdalena Szmajdych, Cracow University of Technology, Poland
Jie Tao, Karlsruhe Institute of Technology, Germany
Pietro Terna, University of Torino, Italy
Peter Trkman, University of Ljubljana, Slovenia
Klaus Troitzsch, University of Koblenz-Landau, Germany
Christopher Tubb, University of South Wales, United Kingdom
Jaroslav Vitku, Czech Technical University in Prague, Czech Republic
Rune Volden, Ulstein Power & Control AS, Norway
Eva Volna, University of Ostrava, Czech Republic
Wei Wang, Beihang University, China
Ivan Yatchev, Technical University Sofia, Bulgaria
Michael Zaggi, Technical University of Munich, Germany
Armin Zimmermann, TU Ilmenau, Germany
Marcelo Zottolo, Lee Memorial Hospital, USA

PREFACE

“If you can’t grow it, you can’t understand it”. This motto summarises at best the idea that only by modelling and simulating a system’s behaviour, we can understand the complexity of the natural, organizational or social reality which we live in.

This is what all the papers included in this collection have in common, independently of their field and domain. In these turbulent times, in which the interdependence of different systems has dramatically grown due to radical, global scale technological innovation, simulation models can help us to improve our understanding of the complex mechanisms that preside over nature, technology, markets and societies, also improving our capacity of managing and finding solutions to complex problems.

The 28th edition of the European Conference on Modelling and Simulation, this year hosted by the University of Brescia, confirms the leading role of this forum in showing the cutting-edge frontier of research in a variety of fields, from engineering and computer sciences to economics and social sciences. It also shows that a trans-disciplinary convergence of different fields towards a common modelling, experimentalist attitude has now strong roots in Europe and worldwide. This can stimulate synergies, complementarities and collaboration that can extend our imagination and creativity.

We strongly believe that crossing the boundaries between fields and disciplines, which are typically compartmentalised and fragmented in academic and research institutions, is fundamental to fuel creativity, innovation and unconventional thinking. It must be said that this is one of the most important roles that ECMS conferences play worldwide and the real secret of their longstanding success.

We hope that the excellent quality of the contributions here included, also guaranteed by the time, dedication and commitment by an impressive number of experts forming the programme committees of the various tracks, will stimulate future directions of research capable of challenging our understanding of nature, technology, organizations, markets and society. We hope that this collection will also promote trans-disciplinary, cross-sectorial collaboration capable of making a real difference.

Although we did our best for the success of this edition, we are sure that next time will be even better!

The conference and programme chairs

Flaminio Squazzoni, Fabio Baronio, Claudia Archetti and Marco Castellani
(University of Brescia)

TABLE OF CONTENTS

Plenary Talks - Abstracts

Preventing Collapse Of Financial Networks Through Systemic Risk Taxes - Answers From Agent Based Models	
<i>Stefan Thurner</i>	5
Patterns, Protocols, And Predictions: Agent-Based Modelling As A Multi-Scope For Analysing Complex Systems	
<i>Volker Grimm</i>	6
Numerical Simulation Of Surface Gravity Waves	
<i>Miguel Onorato</i>	7

Agent-Based Simulation

Use Of Agent Based Modeling To Simulate Complex Ecological Systems In Contexts With Poor Information; The Case Of The Winton Wetlands In Victoria, Australia	
<i>Luisa Perez-Mujica, Terry Bossomaier, Roderick Duncan, Max C. Finlayson</i>	11
Simulating Daily Mobility In Luxembourg Using Multi-Agent Based System	
<i>Hedi Ayed, Benjamin Gateau, Djamel Khadraoui</i>	18
Agent-Based Control Framework In Jade	
<i>Franco Cicirelli, Libero Nigro, Francesco Pupo</i>	25
Simulating Social Networks In Social Marketing	
<i>Roderick Duncan, Luisa Perez-Mujica, Terry Bossomaier</i>	32

Simulation of Complex Systems and Methodologies

Simulation Of A Muon Based Monitoring System	
<i>Germano Bonomi, Antonietta Donzella, M. Subieta, Aldo Zenoni</i>	41
Towards An Executable Sociotechnical Model For Product Development And Engineering Systems	
<i>Axel Hahn, Juergen Geuter</i>	47

Simulation, Experimental Science and Engineering in Maritime Operations

A Hardware-In-The-Loop Simulator For Offshore Machinery Control System Testing

Johnny Aarseth, Alf Helge Lien, Oyvind Bunes, Yingguang Chu, Vilmar AEsøy57

A Ship Motion Short Term Time Domain Simulator And Its Application To Costa Concordia Emergency Manoeuvres Just Before The January 2012 Accident

Paolo Neri, Mario Piccinelli, Paolo Gubian, Bruno Neri.....64

Interoperability In Co-Simulations Of Maritime Systems

Christoph Dibbern, Axel Hahn, Soeren Schweigert71

Simulation and Visualization for Training and Education

Recycling A Discarded Robotic Arm For Automation Engineering Education

Filippo Sanfilippo, Ottar L. Osen, Saleh Alaliyat81

Modelling And Simulation Of An Offshore Hydraulic Crane

Yingguang Chu, Vilmar AEsøy, Houxiang Zhang, Oyvind Bunes87

A Real-Time UAV INSAR Raw Signal Simulator For HWIL Simulation System

Wei Li, Houxiang Zhang, Hans Petter Hildre94

JIOP: A Java Intelligent Optimisation And Machine Learning Framework

Lars I. Hatledal, Filippo Sanfilippo, Houxiang Zhang101

Enhancing Undergraduate Research And Learning Methods On Real-Time Processes By Cooperating With Maritime Industries

Webjorn Rekdalsbakken, Filippo Sanfilippo.....108

Ballast Water Analysis And Heat Treatment Using Waste Heat Recovery Systems On Board Ships

Yanran Cao, Vilmar AEsøy, Anne Stene.....115

Simulation Challenges In The Development Process Of A Complex Product: Design Of Virtual Electric Sports Car

Eszter Varga, Attila Piros, Balazs Vidovics122

Electrical and Electromechanical Engineering

Determination Of An Optimal Shape Of Rotor For The Synchronous Reactive Frequency Doubler

Aleksandrs Mesnajevs, Elena Ketnere 131

Modelling Of Broadband Electric Field Propagation In Nonlinear Dielectric Media

Matteo Conforti, Fabio Baronio 136

On The Processing Of The Recorded Data For The SF6 Circuit Breakers From The Transformation Substation 110/20/6kV Craiova South

Maria Brojboiu, Virginia Ivanov, Andrei Savescu 142

Design, Simulation And Testing Of Planar Spiral Coils For The Time-Gated Interrogation Of Quartz Resonator Sensors

Mohamad Farran, Marco Bau, Daniele Modotto, Marco Ferrari, Vittorio Ferrari 147

Applications Of The Graph Theory For Optimization In Manufacturing Environment Of The Electrical Equipments

Virginia Ivanov, Maria Brojboiu, Sergiu Ivanov 153

Current Control Of A VSI-FED Induction Machine By Predictive Technique

Sergiu Ivanov, Vladimir Rasvan, Eugen Bobasu, Dan Popescu, Florin Stinga 159

Modeling And Control Of Berry Phase In Quantum Dots

Sanjay Prabhakar, Roderick Melnik, Ali Sebetci 166

Efficient Modeling Of Graphene Based Optical Devices

Costantino De Angelis, Andrea Locatelli 171

Simulation in Industry, Business and Services

A Simulation Study: The Business Value Of E-Business For A Maintenance Provider

Orit Raphaeli, Liron Rosenfeld, Lior Fink, Sigal Berman..... 179

Strategic Information Systems Planning As A Dynamic Capability: Insights From An Agent-Based Simulation Study

Daniel Fuerstenau, Johannes Schinzel, Catherine Cleophas..... 185

Extended Neonatal Metabolic Screening By Tandem Mass Spectrometry: Models And Simulation Of Alternative Management Solutions

Arturo Liguori, Giorgio Romanin-Jacur..... 193

The Influence Of Changing Environment For Path Dependence In Hierarchical Organizations

Arne Petermann, Alexander Simon..... 202

HCCM - A Control World View For Health Care Discrete Event Simulation

*Nikolaus Furian, Michael O'Sullivan, Cameron Walker,
Siegfried Voessner*..... 206

Towards A Simulation Model Of Partner-Specific Absorptive Capacity As A Path Dependent Self-Reinforcing Mechanism In B2B Relationships

Tobias Grossmann, Arne Petermann..... 214

Simulation Of Scheduling And Cost Effectiveness Of Nurses Using Domain Transformation Method

Geetha Baskaran, Andrzej Bargiela, Rong Qu 226

Emergency Department: A General Adaptable Simulation Model Implemented In Arena

Arturo Liguori, Giorgio Romanin-Jacur..... 235

Modelling, Simulation and Control of Technological Processes

Nonlinear Control Of A Shell And Tube Heat Exchanger <i>Petr Dostal, Vladimir Bobal, Jiri Vojtesek</i>	247
Digital Linear Quadratic Smith Predictor <i>Vladimir Bobal, Marek Kubalcik, Petr Dostal, Stanislav Talas</i>	254
Simulation Of Multivariable Continuous-Time Decoupling Control <i>Marek Kubalcik, Vladimir Bobal</i>	261
Use of Dynamic Matrix Control in Simulation of Heat System <i>Stanislav Talas, Vladimir Bobal, Adam Krhovjak</i>	267
Multivariable Adaptive Control Of Two Funnel Liquid Tanks In Series <i>Adam Krhovjak, Petr Dostal, Stanislav Talas</i>	273
Modeling Of Alcohol Fermentation In Brewing – Carbonyl Compounds Synthesis And Reduction <i>Vessela Naydenova, Vasil Iliev, Maria Kaneva, Georgi Kostov Petia Koprinkova-Hristova, Silviya Popova</i>	279
Robust Process Control With Saturated Control Input <i>Frantisek Gazdos, Jiri Marholt</i>	285
Computational Model For Spray Quenching Of A Heavy Forging <i>Mahdi Soltani, Annalisa Pola, Giovina Marina La Vecchia</i>	292
Modelling And Simulation Of Water Tank <i>Jiri Vojtesek, Petr Dostal, Martin Maslan</i>	297
The Effect Of Initial Estimated Points On Objective Functions For Optimization <i>Mahdi Soltani, Annalisa Pola, Qiang Xu</i>	304
Estimation Of The Dynamic Effect In The Lifting Operations Of A Boom Crane <i>Luigi Solazzi, Giovanni Incerti, Candida Petrogalli</i>	309

Simulation of Intelligent Systems

Towards Evolutionary Deep Neural Networks

*Tomas H. Maul, Andrzej Bargiela, Siang-Yew Chong,
Abdullahi S. Adamu*319

Model Of Intellectual Visualization Of Geoinformation Service

*Stanislav L. Belyakov, Alexander V. Bozhenyuk,
Marina L. Belykova, Igor N. Rozenberg*326

Advantages Of Using Memetic Algorithms In The N-Person Iterated Prisoner's Dilemma Game

*Tamara Alvarez, Miguel Loureiro, Jose Covelo, Ana Peleteiro,
Aleksander Byrski, Juan C. Burguillo*333

A Comparative Study To Evolutionary Algorithms

Eva Volna, Martin Kotyrba340

Comparison Of Modern Clustering Algorithms For Two-Dimensional Data

Martin Kotyrba, Eva Volna, Zuzana Kominkova Oplatkova.....346

Reusable Reinforcement Learning for Modular Self Motivated Agents

Jaroslav Vitku, Pavel Nahodil352

Routing And Communication Path Mapping In VANETS

Nnamdi Anyameluhor, Evtim Peytchev.....359

Modelling Retinal Feature Detection With Deep Belief Networks In A Simulated Environment

Diana Turcsany, Andrzej Bargiela, Tomas Maul.....364

Performance Comparison Of Evolutionary Techniques Enhanced By Lozi Chaotic Map In The Task Of Reactor Geometry Optimization

Michal Pluhacek, Roman Senkerik, Ivan Zelinka, Donald Davendra371

Analysis Of EEG Signal For Using In Biometrical Classification

Roman Zak, Jaromir Svejda, Roman Senkerik, Roman Jasek377

Using Artificial Neural Network For The Kick Techniques Classification – An Initial Study

*Dora Lapkova, Michal Pluhacek, Zuzana Kominkova Oplatkova,
Milan Adamek*.....382

Pseudo Neural Networks For Iris Data Classification

Zuzana Kominkova Oplatkova, Roman Senkerik, Ales Kominek.....387

Simulation Of The Differential Evolution Performance Dependency On Switching Of The Driving Chaotic Systems	
<i>Roman Senkerik, Michal Pluhacek, Donald Davendra, Ivan Zelinka, Zuzana Kominkova Oplatkova</i>	393

Modelling and Simulation in Robotic Applications

Gyroscopic Precession In Motion Modelling Of Ball-Shaped Robots	
<i>Tomi Ylikorpi, Pekka Forsman, Aarne Halme</i>	401
Unified Representation Of Decoupled Dynamic Models For Pendulum-Driven Ball-Shaped Robots	
<i>Tomi Ylikorpi, Pekka Forsman, Aarne Halme, Jari Saarinen</i>	411
Integrating Simulation With Robotic Learning From Demonstration	
<i>Anat Hershkovitz Cohen, Sigal Berman</i>	421

Discrete Event Modelling and Simulation in Logistics, Transport and Supply Chain Management

Improving The Distribution Planning Process In The Food&Beverage Industry: An Empirical Case Study	
<i>Andrea Bacchetti, Massimo Zanardini</i>	431
A Timed Petri Net Model For The Quay Crane Scheduling Problem	
<i>Roberto Trunfio</i>	441
Determining Transportation Mode Choice To Minimize Distribution Cost: Direct Shipping, Transit Point And 2-Routing	
<i>Luca Bertazzi, Jeffrey Ohlmann</i>	448
Process Modelling And Simulation For Medication-Use Process	
<i>Johan Royer, Michelle Chabrol, Jean-Luc Paris</i>	454
A Scor Based Analysis Of Simulation In Supply Chain Management	
<i>Wolfgang Kersten, Muhammad Amad Saeed</i>	461

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems Special Session

Implementation Of The Genetic Algorithm By Means Of CUDA Technology Involved In Travelling Salesman Problem

*Anna Plichta, Tomasz Gaciarz, Bartosz Baranowski, Szymon Szominski.....*475

Workload Characterization Of Multithreaded Applications On Multicore Architectures

*Davide Cerotti, Marco Gribaudo, Mauro Iacono, Pietro Piazzolla.....*480

Dynamic Factory - New Possibilities For Factory Design Pattern

*Dawid R. Ireno.....*487

Memetic Computing In Selected Agent-Based Evolutionary Systems

*Aleksander Byrski, Marek Kisiel-Dorohinicki.....*495

Optimal Pump Scheduling By NLP For Large Scale Water Transmission System

*Jacek Blaszczyk, Krzysztof Malinowski, Alnoor Allidina.....*501

Hybrid CPU/GPU Platform For High Performance Computing

*Michal Marks, Ewa Niewiadomska-Szynkiewicz.....*508

Using Artificial Neural Network For Monitoring And Supporting The Grid Scheduler Performance

*Daniel Grzonka, Joanna Kolodziej, Jie Tao.....*515

Hybrid Architecture For Simulation Of Blood Flow With Foreign Bodies

*Lukasz Faber, Krzysztof Boryczko, Marek Kisiel-Dorohinicki.....*523

Realistic Mobility Simulator For Smart Traffic Systems And Applications

*Cosmin-Stefan Stoica, Ciprian Dobre, Florin Pop.....*530

Research and Use of Multiformalism Modeling Methods Special Session

Multi-Formalism Modeling For Evaluating The Effect Of Cyber Exploits

*Alexander H. Levis, Bahram Yousefi.....*541

Probability and Statistical Methods for Modelling and Simulation of High Performance Information Systems Special Session

Analysis Of A FCFS Queue With Two Types Of Customers And Order-Dependent Service Times

Bert Reveil, Dieter Claeys, Tom Maertens, Joris Walraevens, Herwig Bruneel551

Joint Stationary Distribution Of Queues In Homogenous M|M|3 Queue With Resequencing

Ilaria Caraccio, Alexander V. Pechinkin, Rostislav V. Razumchik558

Generation Of Probability Measures With The Given Specification Of The Smallest Bans

Alexander A. Grusho, Nick A. Grusho, Elena E. Timonina565

On The Development Of An Information Technology For Plasma Turbulence Research

Andrey Gorshenin, Victor Korolev, Dmitry Malakhov, Nina Skvortsova, Sergey Ya. Shorgin, Victor Kuzmin.....570

On Truncations For SZK Model

Alexander Zeifman, Yakov Satin, Galina Shilova, Victor Korolev, Vladimir Bening, Sergey Ya. Shorgin.....577

On Convergence Of The Distributions Of Random Sums And Statistics Constructed From Samples With Random Sizes To Exponential Power Laws

Victor Korolev, Maria E. Grigoryeva, Alexander Zeifman.....583

Transfer Theorem Concerning Asymptotic Expansions For The Distribution Functions Of Statistics Based On Samples With Random Sizes

Vladimir Bening, Vladislav A. Savushkin, Egor I. Shunkov, Alexander Zeifman, Victor Korolev590

Variance-Mean Mixtures As Asymptotic Approximations

Victor Korolev, Vladimir Bening, Andrey Gorshenin, Maria Grigoryeva, Alexander Zeifman596

Analytical Modelling And Simulation For Performance Evaluation Of SIP Server With Hysteretic Overload Control

Konstantin E. Samouylov, Pavel O. Abaev, Yuliya V. Gaidamaka, Alexander V. Pechinkin, Rostislav V. Razumchik603

Simulation and Optimization

Regulation Of The Input Flow Of Supply Chains To Optimize The Production

Ciro D'Apice, Carmine De Nicola, Rosanna Manzo613

Seasonal Trends In Supply Chains

Hans-Peter Barbey620

The Tyche And Safe Models: Comparing Two Military Force Structure Analysis Simulations

Cheryl Eisler, Slawomir Wesolkowski, Daniel T. Wojtaszek625

Modeling Optimal Allocation Centers In GIS By Fuzzy Base Set Of Fuzzy Interval Graph

Leonid S. Bershtein, Alexander V. Bozhenyuk, Stanislav L. Belyakov, Igor N. Rozenberg.....633

Approaches To Run Simulations Of Business Processes In A Grid Computing Network

Christian Mueller.....639

Optimisation Of Boids Swarm Model Based On Genetic Algorithm And Particle Swarm Optimisation Algorithm (Comparative Study)

Saleh Alaliyat, Harald Yndestad, Filippo Sanfilippo643

Modelling Of Photovoltaic Energy Generation Systems

Pekka Ruuska, Antti Aikala, Robert Weiss651

Genetic Algorithm With Simulation For Scheduling Of A Flow Shop With Simultaneously Loaded Stations

Frank Herrmann.....657

MapReduce Based Experimental Frame for Parallel and Distributed Simulation Using Hadoop Platform

Byeong Soo Kim, Sun Ju Lee, Tag Gon Kim, Hae Sang Song.....664

Deriving A Mathematical Model Of A Paint Shop From Data Analysis

Thomas Husslein, Christian Danner, Markus Seidl, Joerg Breidbach, Wolfgang Lauf.....670

Multi-Criteria Approach For Emergency Service Orders In Electric Utilities

Vinicius Jacques Garcia, Daniel Pinheiro Bernardon, Alzenira Abaide, Julio Fonini.....676

An Improved Receding Horizon Genetic Algorithm For The Tug Fleet Optimisation Problem	
<i>Robin T. Bye, Hans G. Schaathun</i>	682

The Value Of Integration In Logistics	
<i>Claudia Archetti, M. Grazia Speranza</i>	691

Finance, Economics and Social Science

Petri Nets As Tools For Policy Analysis: The Example Of Smoking Bans In Public Places	
<i>Georg P. Mueller</i>	701

The Association Between Group Size And Communicational Complexity According To Conceptual Agreement Theory	
<i>Enrique Canessa, Carlos Barra, Sergio E. Chaigneau, Ariel Quezada</i>	709

The Dangers Of Ethnocentrism	
<i>Giangiacoimo Bravo</i>	718

Do Editors Have A Silver Bullet? An Agent-Based Model Of Peer Review	
<i>Juan Bautista Cabota, Francisco Grimaldo, Flaminio Squazzoni</i>	725

Learning not to Trade: On Scarcity, Emergence and Failure of Markets	
<i>Oezge Dilaver</i>	732

Degree Variance And Emotional Strategies Catalyze Cooperation In Dynamic Signed Networks	
<i>Simone Righi, Karoly Takacs</i>	738

A Social Interaction Model For Crime Hot Spots	
<i>Evan C. Haskell</i>	745

The Definition Of Stress Situations And Their Prediction Using Liquidity In The Framework Of The EMIR Regulation	
<i>Barbara Doemoetoer, Kata Varadi</i>	752

Path Dependency In Investment Strategies – A Simulation Based Illustration	
<i>Agnes Vidovics-Dancs, Peter Juhasz, Janos Szaz</i>	758

Modelling The Collapse Of A Criminal Network	
<i>Martin Neumann, Ulf Lotzmann</i>	765

Policy Modelling

Integrating Optimisation And Agent-Based Modelling

Peter George Johnson, Tina Balke, Lars Kotthoff.....775

Towards An Agent-Based Model On Co-Diffusion Of Technology And Behavior: A Review

Thorben Jensen, Emile Chappin.....782

Scenario Analysis And Optimization Approach In Air Quality Planning: A Case Study In Northern Italy

*Claudio Carnevale, Giovanna Finzi, Anna Pederzoli, Enrico Turrini,
Marialuisa Volta*789

Tactical Versus Operational Discrete Event Simulation: A Breast Screening Case Study

Andrea Lodi, Paolo Tubertini, Roberto Grilli, Francesca Senese.....796

Author Index803

ECMS 2014

SCIENTIFIC PROGRAM

Plenary Talks

**Preventing Collapse Of Financial Networks
Through Systemic Risk Taxes
- Answers From Agent Based Models**

-abstract-

Stefan Thurner

Medical University of Vienna

Austria

Financial markets are exposed to systemic risk (SR), the risk that the system ceases to function and collapses. Since recently, it is possible to quantify SR in terms of underlying financial (multiplex) networks where nodes represent financial institutions, and links capture financial contracts such as loans, credits, or derivatives. We show that it is possible to quantify in real data the SR of individual transactions in a financial network. We propose a tax on individual transactions that is proportional to their contribution to the overall SR. If a transaction does not increase SR, it is tax free.

We demonstrate with an agent based model (CRISIS macro-financial model) that the proposed Systemic Risk Tax (SRT) leads to a self-organized re-structuring of financial networks that are practically free of SR.

ABM predictions agree remarkably well with the empirical data and can be used to understand the relation of credit risk and SR.

Patterns, Protocols, And Predictions: Agent-Based Modelling As A Multi-Scope For Analysing Complex Systems

-abstract-

Volker Grimm

Helmholtz Centre for Environmental Research

—

UFZ, Leipzig, Germany

Systems comprised of decision-making agents such as ecosystems or financial markets are complex. Nevertheless, they generate patterns in structure and dynamics which can be observed at different hierarchical levels and scales. Modellers, though, often focus on only one pattern, which usually is not sufficient to select among alternative model formulations. Therefore, pattern-oriented modelling (POM) has been developed as a general strategy for using multiple observed patterns for the multi-criteria design, selection and calibration of models of complex systems. Instead of using models as a ‘micro-scope’ focussing on individuals, or ‘macro-scope’ focussing on systems dynamics, POM uses agent-based models as ‘multi-scopes’ to capture the interaction between the whole system and its building blocks.

I will present examples from ecology and other domains. I will demonstrate that models developed according to POM usually have a high level of structural realism, i.e. a high chance of capturing the internal organization good enough to make robust predictions of system responses to new conditions. Still, POM models are often tied to specific systems and observations.

To proceed to a more general theory of agent-based complex systems, the ODD protocol for communicating agent-based models can be used to systematically relate the structure and processes of models to broad classes of patterns, or stylized facts, observed in different systems.

Numerical Simulation Of Surface Gravity Waves

-abstract-

Miguel Onorato

University of Turin

Italy

The dynamics of surface gravity waves, i.e. waves at the interface between water and air, is governed by the Navier-Stokes equations that account for the conservation of momentum and mass of a small but macroscopic element of fluid. Boundary conditions at the free surface are required in order to describe the dynamics of the interface.

The numerical simulations of the Navier-Stokes equations is in general not an easy task, especially if the Reynolds number is large enough and turbulence takes place.

In the specific case of surface gravity waves, the computation is even more complicated by the fact that two fluids (air and water) are part of the domain. Moreover, waves are generated by a turbulent wind and waves may go through a breaking process in which air is entrapped in water forming bubbles.

In the talk I will present the state of the art of the simulations of ocean waves and discuss some recent results obtained using

- 1) a level set method (in collaboration with A. Iafrati) and
- 2) boundary-fitted approach (in collaboration with F. Zonta).

Agent-Based Simulation

Use of Agent Based Modeling to simulate complex ecological systems in contexts with poor information; the case of the Winton Wetlands in Victoria, Australia

Luisa Perez-Mujica
School of Environmental Sciences
Charles Sturt University
Elizabeth Mitchel Drive, Albury, 2640
Australia
E-mail: lperezmuja@csu.edu.au

Terry Bossomaier
Centre for Research in Complex
Systems, Charles Sturt University
Panorama Avenue, Bathurs, 2795
Australia
E-mail: tbossomaier@csu.edu.au

Roderick Duncan
School of Accounting and Balance
Charles Sturt University
Panorama Avenue, Bathurs, 2795
Australia
E-mail: rduncan@csu.edu.au

Max C Finlayson
Institute of Land, Water and Society
Charles Sturt University
Elizabeth Mitchell Drive, Albury, 2640
Australia
E-mail: mfinlayson@csu.edu.au

ABSTRACT

There have been numerous frameworks and approaches developed for the study of complex ecological systems. For the most part, these approaches require extensive amounts of information. The aim of this paper is use Agent-Based Models to simulate complex ecological systems and examine overarching trends, using the Winton Wetlands in Victoria, Australia as a case study. The study showed that even though there are gaps of information about the functioning of this particular site, this type of framework could increase the understanding of complex ecological systems and assist in decision-making processes.

INTRODUCTION

In the last decades, ecological systems, as complex systems, have been gaining relevance in scientific, management and policy spheres (An 2012; Young et al. 2006). The study of their modeling has lead to the development of several tools in order to better understand the interactions within ecological systems when many elements are involved (Ford 2010; Grimm and Railsback 2005), and in many cases, in an attempt to further incorporate them into larger systems i.e. socio-ecological systems (Binder et al. 2013; Perez-Mujica et al. 2013).

The study and management of ecological systems, and wetlands in particular, requires the development of integrated management plans covering all the aspects of the wetlands and their relationships with their catchments (Finlayson et al. 2005).

Lack of information, uncertainty, non-linear relations and emergence of system's behaviors are some the common features that characterize the analysis of complex systems and require to be taken into account during the modeling process (Ladyman et al. 2013;

Sterman 2001).

Top-down simulation frameworks and aggregational models, such as differential equations have been most commonly used to model complex ecological systems (Ford 2010). Top-down frameworks require thinking in terms of global structural dependencies and the behavior of the system is defined in the onset. (Bonabeau 2002; Scholl 2001). However, for some systems, we might know very little about how things affect each other at the aggregate level, or what is the global consequence of operation, but we might have some perception of how individual participants of the process behave. In instances like this one, Agent Based Modeling can be useful.

Agent-Based Modeling (ABM) has been gaining importance as a tool to understand complex systems, (Epstein 2006; Gilbert 2008; Gilbert and Bankes 2002). As a bottom-up approach the behavior of individual agents accounts for lack of knowledge of the general structure of the system, as well as for the adaptation of the agents in response to the changes in the environment (Bonabeau 2002). In this regard, ABMs can be used to look for trends and patterns at a level where there is just sufficient data to capture the dynamics of the system (Epstein 2006).

Authors of ABM, or Individual Based Modeling in Ecology, define agents as elements with autonomous decisions that help them achieve their goals (survival, reproduction, etc.) and adapting the decisions to the rapidly changing environment (Grimm and Railsback, 2005; Grimm et al. 2005).

ABMs have been used in the context of ecological systems to describe the structure and interactions in complex ecological systems, from development and growth of beech forests to selection of habitat by trout (Grimm and Railsback 2005; Grimm et al. 2005). In addition, ABMs have also been used in the context of

environmental management to simulate and model of the effect of human decisions on the environment with the aim of evaluating policies and support decision making processes (Bennett and Tang 2006; Tang and Bennett 2010) but they are still far from being used extensively for ecological systems.

Wetlands, as complex systems, are comprised of a network of interactions among different biotic and abiotic elements, which make complete representation extremely difficult (Powell et al. 2008). The Winton Wetlands is no exception. Over the years, the site has been subjected to different hydrological and ecological regimes. It was transformed from a network of wetlands to agricultural land, followed by an irrigation lake and finally turned back into natural wetlands during a process of restoration (Winton Wetlands Committee of Management Inc 2011). These dramatic changes of land use are associated with different levels of information. There is little or close to no information about the state of the network of wetlands prior to the establishment of European settlers and during the first two years after the creation of the irrigation lake. The main body of literature available was collected during the 1990s when it was still a lake and during the decommissioning process of the lake. Although the current Committee of Management of the restoration project has directed certain monitoring programs, there is little knowledge about the behavior of the system, which would aid in the decision-making process

This paper is part of a larger project about the Winton Wetlands, focusing on the development of a systemic approach to study complex socio-ecological systems in the context of sustainability (Perez-Mujica et al. 2013; Perez-Mujica et al. 2014) and its validation with the participants is currently in process.

Drawing on the best available information, including quantitative modeling and qualitative “expert opinion”, the aim of this paper is use Agent-Based Models to simulate complex ecological systems and examine overarching trends, using the Winton Wetlands in Victoria, Australia as a case study. The ABM will simulate the ecological interactions resulting from different climatic and hydrological scenarios for two distinct areas inside the site: the wetlands and the woodland.

The aim is not to provide an accurate prediction of the future of the site but to increase our understanding of about the emergent behavior, i.e. the dynamics at the level of the whole system. The use of ABM in systems like this could help improve the level of understanding of complex ecological systems with information gaps, help in the decision making process and guide future research and monitoring plans.

METHODS

Study area

The Winton Wetlands is a transformed wetland located in North East Victoria in Australia, approximately

200km north from Melbourne. After its decommissioning as a lake, the state government has implemented a restoration project to return, as much as possible the Winton Wetlands to its original state (Goulburn Broken Catchment Management Authority 2012).

Data collection and analysis

Data was collected in the form of 13 semi-structured interviews of the Scientific and Technical Advisory Group of the Winton Wetlands, as well as wetland ecologists and conservation employees of the Catchment Management Authority that have the Winton Wetlands within their jurisdiction. Interviews were transcribed and coded for themes in Nvivo, using the Actors, Factors, Sectors Framework proposed by Kok and colleagues (2006) to characterize socio-ecological issues (Perez-Mujica et al. 2013). Where possible, official documents about the ecology of the Winton Wetlands were used to complement the information used in the model, e.g. the Restoration and Monitoring Strategic Plan (Winton Wetlands Committee of Management Inc 2011).

The elements and interactions were represented in a conceptual model and then converted to an ABM. The implementation language was Netlogo (Wilensky 1999).

The Model

Based on the interviews and the documentation, a conceptual model of the Winton Wetlands was established (Figure 1). The general elements and interactions presented in the model represent causal interactions established by the participants and recorded as relationship themes in Nvivo. The basic structure of the study area is divided into two main areas of the site: the wetland and the woodland. The wetland portion is comprised of the area of the site that can be flooded. The woodland portion is the permanently dry part of the site.

Ephemeral wetlands, such as the Winton Wetlands are complex systems whose ecological integrity is intimately linked with natural hydrological and climatic cycles, which results in the increase and decrease of the quantity of water in the wetland (Powell et al. 2008). Most of the other elements of the system depend upon it. When there is water in the system (via rainfall and water inflows from the catchment), the quantity of water in the wetland, allows fish populations to enter the system from river feeds and woodland vegetation to increase. When there is water in the wetland, species of migratory birds also visit the wetland and they leave when there is no more water.

However there are some threats to the ecological function of the site. These can be in the form of introduced pests, such as exotic fish and feral animals, as well as exotic plants.

To specify the ABM, the different variables were extracted from the conceptual model (Table 1). Because there are two distinct portions of the site, there are two

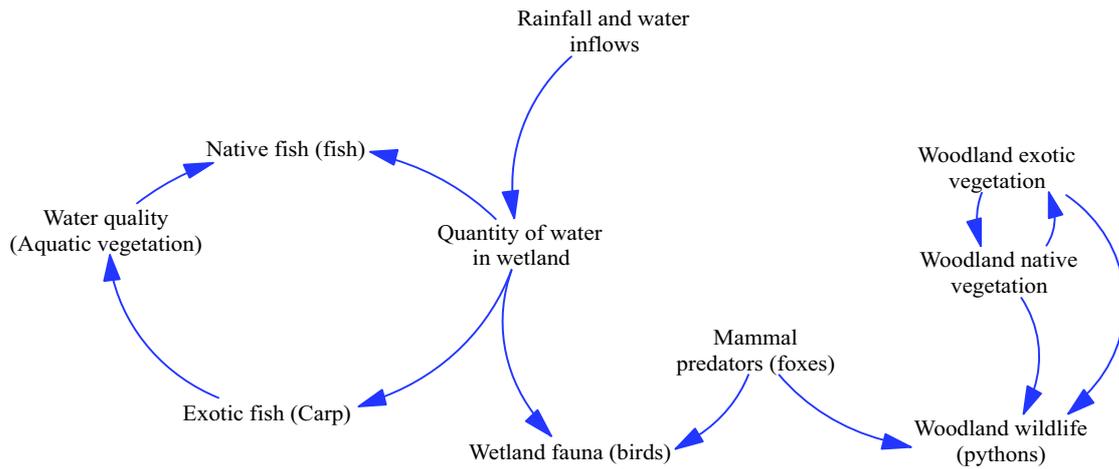


Figure 1: conceptual map of the Winton Wetlands ecological system

groups of behavioral rules; one for the wetland and one for the woodland (Figure 2, Figure 3).

As mentioned previously, rainfall and the subsequent amount of water in the site determine many of the interactions within the system. In the case of the wetlands, the amount of rainfall from year to year is random, but there are two distinct stages; a drought stage (in most years the wetland retains some water but can at times dry out completely) and decadal flooding that refills the entire wetland. Based on this, there are four main behaviors for the wetland (Figure 2):

1. If there is no rainfall and it corresponds to a period of drought where the wetlands are completely dry, all the carp and native fish die, and the wetland becomes a seed bank for new aquatic plants.
2. If there is a big flood, the wetland is refilled and native and exotic fish enter the system via streams from the surrounding catchment.
3. If there is rainfall but it doesn't correspond to a big flood, there is a mosaic of areas with and without water. In the areas with water, carp, which churn mud and disturb aquatic vegetation, decrease the quality of water, which in turn affects the population of native fish.

On the other hand, in the woodland part of the reserve (Figure 3) there is a competitive interaction between native woodland vegetation and exotic vegetation. Through wind, both types of vegetation disperse their seeds to colonize new spaces but the exotic species are far more successful and vigorous than native species. Nevertheless, through competitive exclusion, once one type of vegetation types establishes, the other cannot establish itself anymore.

Table 1. Variables of the Winton Wetlands ABM

Variable/ type	Description
Wetland	Wetland portion of the site
Carp/agent	Exotic fish that enter the wetland every big flood
Fish/agent	Native fish that enter the wetland every big flood
Birds/agent	Terrestrial fauna associated with the wetlands
Water quality/patch	Level of health of the aquatic vegetation of the wetland. It is a proxy of the ecological function that the wetland provides to aquatic biota.
Shore/patch	Dry portion of the wetland
Woodland	Woodland portion of the site
Water quantity/patch	Amount of water in the wetland
Pythons/agent	Native woodland wildlife
Native plants/patch	Proportion of native woodland flora
Exotic plants/patch	Proportion of exotic woodland flora

Finally, there is a predatory interaction in both portions of the reserve, wetland and woodland, involving animal pests (in this case, foxes and feral cats). They predate on terrestrial fauna associated with the wetlands, such as migratory birds, as well as on woodland wildlife, such as carpet pythons. In the case of the predation on birds, this stops once the birds leave the wetland when it is completely dry.

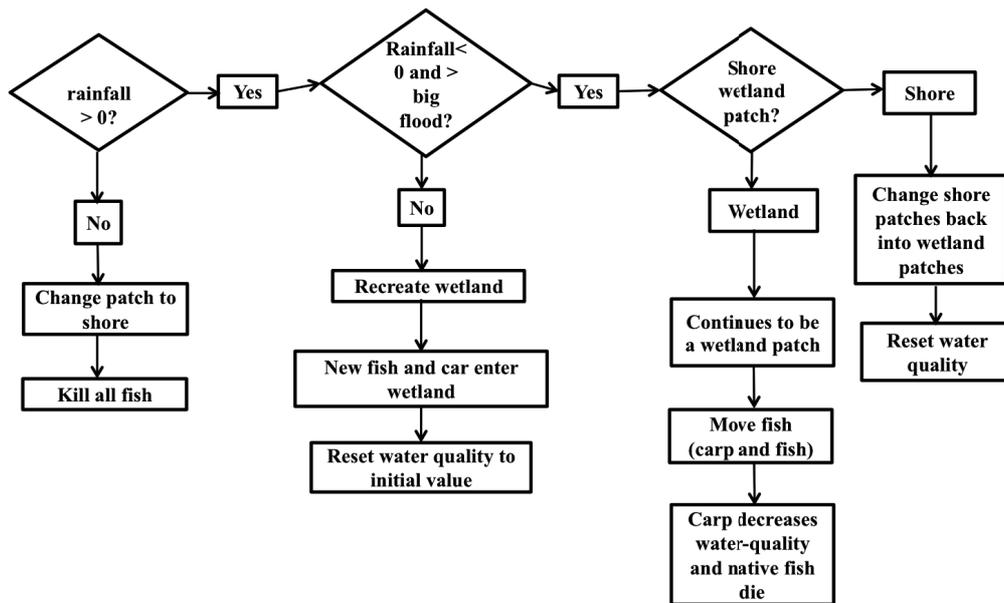


Figure 2: Behavioral rules for the dynamics of the wetland

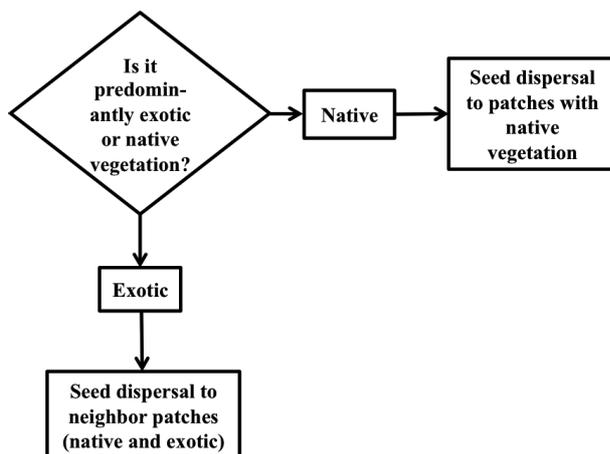


Figure 3: Behavioral rules for the dynamics of the woodland vegetation

RESULTS

For the wetlands, two different scenarios were simulated based on the important elements and interactions gathered from the interviews. Most participants mentioned that the water quality and quantity in the wetland were the two most important elements in the wetland system.

Firstly, the quantity of water in the system responds to the amount of annual rainfall and the presence of big flooding events.

In general, it was stated that the rainfall events are unpredictable but that approximately every ten years the wetlands would receive an important influx of water from the nearby rivers and that during the period in between these “big floods” the wetlands would have different levels of water and would eventually be completely dry at some point. Secondly, it was established that the presence of carp, in terms of the disturbance of sediment and subsequent increase in turbidity, were some of the main causes for the decrease of water quality. When the quality of water decreases other native species can no longer survive and the population of native species (fish in this model) decreases.

The difference between scenario (a) and (b) is that the first one has a bigger population of carp in the wetland. Both scenarios have the same amount of fish (native species) and initial quality of water.

When there is an increase in the entrance of carp to the system through the big floods, i.e. there is not an efficient management plan to control the entrance of carp, the population of native fish and water quality decreases (Figure 4a).

When the entry of carp is controlled, by means of fish way traps for example, (cages at the inlet channels of the wetlands), the water quality is more stable and the population of fish does not decrease (Figure 4b).

In both scenarios, the wetlands have periods of complete dryness, in which both the carp and native fish die.

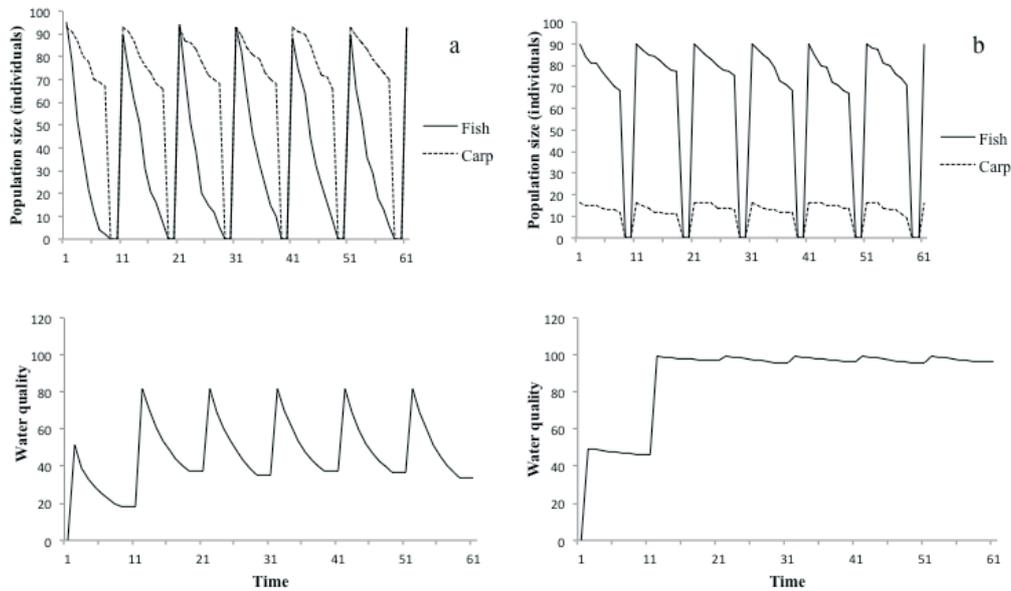


Figure 4: Two scenarios for carp and fish populations and water quality; one with a high initial population of carp (a) and one with low initial population of carp (b)

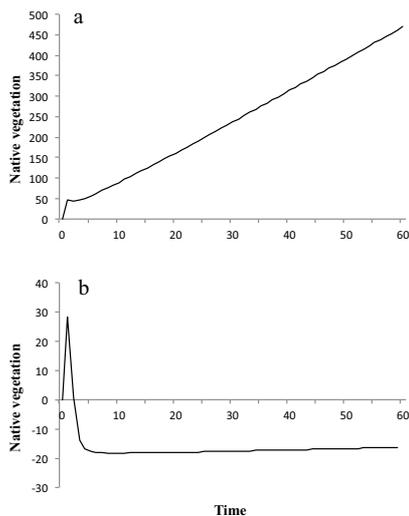


Figure 5: Two scenarios for native vegetation; one with high initial native vegetation (a) and one with low initial native vegetation (b)

When this happens, the wetlands goes through a process of “refreshing” according to a participant expert in wetland restoration in which the wetlands recycle nutrients, among other processes that allow native vegetation and fauna to reestablish once the wetland is refilled.

In this regard participants agreed that even though people normally want to see the wetlands always pretty and full of water, it is in fact important for the wetlands to dry out.

For the woodlands, there are three important elements and several interactions. Firstly, all participants agreed that the biggest problem of the woodland and grassland is the dominance of a few introduced species, which prevent colonization by native species. Without the implementation of management plans that control the spreading of exotic plants in the system, the survival and proliferation of native plants is compromised (Figure 5b).

However, if management plans are successful in controlling these weeds, the ratio of native vegetation to exotic plants increases because native plants are able to colonize (Figure 5a).

Finally, two scenarios were developed to simulate the response of the populations of native woodland fauna (pythons) and associated terrestrial fauna (birds) as a result of the predation of foxes.

Animal pests (foxes and feral cats) are ubiquitous to this region of Australia. A large proportion of reserves in Victoria have problems with foxes. A participant mentioned that the programs of pest control are very important to keep in check the population of animal pests. Although the population of birds and pythons is decreasing due to the predation of foxes, in a scenario where there was a high initial number of foxes (Figure 6b) the rate of decrease was larger than where there was a low initial number of foxes (Figure 6a).

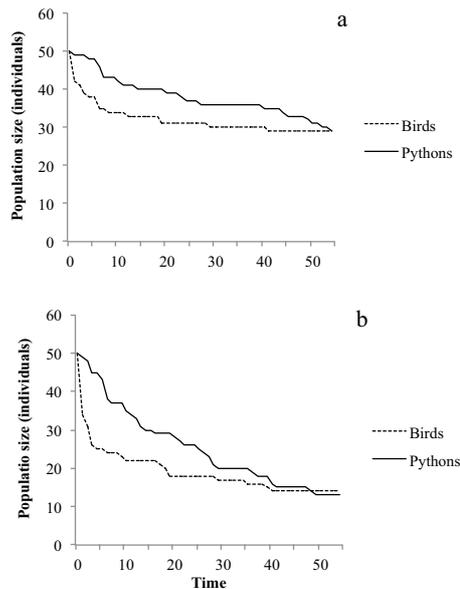


Figure 6: Two scenarios native terrestrial populations; one with high initial number of foxes (a) and one with low initial number of foxes (b)

DISCUSSION

The simulation in the wetlands showed that the population of native fish greatly depended on the population of carp because of the relationship of carp and the decrease of water quality.

However, it is noteworthy that all species of fish, native and exotic also depend on the amount of water in the wetland. So, even if carp managed to get into the system during one of the big flooding periods, once the wetland is completely dried out, all the fish will die and the wetland would refresh. In this regard, the simulation helps to communicate to a wider audience, that is 'non-scientist' about the importance of ephemeral wetlands being wet and dry in different periods.

In addition to "clearing" the wetland of exotic fish and recycling nutrients, during the dry period the wetland is also habitat for some species that require the soil to be drier. A next step of this model would be to incorporate the dynamics of biodiversity found during the drying period of the wetlands with species like wading shore birds, such as gulls.

It can also be seen that even though the wetlands would refresh every big flood, once fewer carp enter the system, the quality of water stays at a similar level until it is entirely dried. This would imply that the body of water would have a high quality of water in the periods between big floods and would be available to sustain an array of native flora and fauna.

Another important element of the system that depends on the quantity of water is the presence and absence of wetland associated fauna (birds in this model). However, the decrease in population size due to the decrease in the amount of water does not necessarily

represent the death of the birds. Birds are mobile animals and they will migrate wherever they find water to feed on the fish. Once the wetland is full again, the birds will migrate back to the wetlands.

Predation by feral animals (foxes in this model) does reflect directly the death of associated fauna. A further step in the model would be to separate the migration away from the wetlands from birds as a result of the decrease on quantity of water and the death of birds as a result of predation from foxes

For the woodlands, the model represents the biggest issues mentioned by the participants was the ratio off native vegetation to exotic vegetation. In the case of the woodlands, participants stated that the management of exotic plants is much more complex than the wetland because "the wetland dries and floods and takes care of itself". In the woodland, things are different, vigorous exotic vegetation will dominate the landscape.

The simulation for the woodlands and the wetlands helps to shed light on the importance of implementing accurate and efficient management plans to control animal pests and weeds. If the carp are never managed by controlling their entry to the wetland, native aquatic species populations will struggle to survive in the system. Something similar occurs for the woodland, control of exotic vegetation and animal pests. Re-vegetation of native species needs to be programmed and implemented in order to restore the ecological function of the woodland. This will possibly never be achieved to the original pre-European settlement state, but closer to a functioning woodland state.

As mentioned before, this model is part of a bigger project. It is still in the process of development and collection of empirical data for more precise parameterization. Future work will investigate the interaction of social and economic subsystems.

AUTHOR BIOGRAPHIES

LUISA PEREZ-MUJICA holds a Bachelor degree in Biology with Honours in Environmental Law and Policy at the UNAM. She worked as a research assistant for the National Laboratory of Sustainability Sciences in Mexico City before moving to Australia where she is now conducting her PhD at Charles Sturt University. Her e-mail address is lperezmuji@csu.edu.au.

PROFESSOR TERRY BOSSOMAIER is the Director of the Centre for Research in Complex Systems of Charles Sturt University at the campus in Bathurst, Australia. Terry is a computational scientist with interests in the theory and applications of complex systems. His email address is tbossomaier@csu.edu.au

DR. RODERICK DUNCAN is a senior lecturer in economics and finance at Charles Sturt University at the Bathurst campus in Australia. He holds a degree in economics and law at the Australian National

University and his doctorate in economics at Stanford University. His e-mail is rduncan@csu.edu.au

PROFESSOR MAX C FINLAYSON is the Director of the Institute of Land, Water and Society at Charles Sturt University in Albury, Australia. He is an internationally renowned wetland ecologist and climate change, and human wellbeing and wetlands. His e-mail is mfinlayson@csu.edu.au

REFERENCES

- An, L. 2012. "Modeling human decisions in coupled human and natural systems: Review of agent-based models". *Ecological Modelling*, 229.
- Bennett, D. A. and W Tang, 2006. "Modelling adaptive, spatially aware, and mobile agents: Elk migration in Yellowstone". *International Journal of Geographical Information Science* 20, No. 9, 1039-1066.
- Binder, C.; J. Hinkel; P. Bots; and C.Pahl-Wostl. 2013). "Comparison of frameworks for analyzing social-ecological systems". *Ecology and Society* 18, No. 4, 26.
- Bonabeau, E. 2002. "Agent-based modeling: Methods and techniques for simulating human systems". *Proceedings of the National Academy of Sciences of the United States of America* 99, Suppl 3, 7280-7287.
- Epstein, J. M. 2006. "Generative social science: Studies in agent-based computational modeling", Princeton University Press.
- Finlayson, C.; M Bellio and J. Lowry. 2005. "A conceptual basis for the wise use of wetlands in northern Australia: linking information needs, integrated analyses, drivers of change and human well-being." *Marine and Freshwater Research* 56, No. 3, 269-277.
- Ford, A. 2010. *Modeling the Environment* (2nd ed.), Island Press, Washington, D.C.
- Gilbert, N. 2008. Agent-based models, Sage.
- Gilbert, N. and S Banks. 2002. "Platforms and methods for agent-based modeling". *Proceedings of the National Academy of Sciences of the United States of America* 99, Suppl 3, 7197-7198.
- Goulburn Broken Catchment Management Authority. 2012. *Lake Mokoan Decommissioning*. Retrieved September 9th, 2012, from http://www.gbcma.vic.gov.au/default.asp?ID=lake_mokoan
- Grimm, V. and S. Railsback. 2005. *Individual-based modeling and ecology*. Princeton University Press.
- Grimm, V.; E Revilla; U. Berger; F. Jeltsch; W Mooij; S. Railsback; H. Thulke; J. Weiner; T. Wiegand and D DeAngelis. 2005. "Pattern-oriented modeling of agent-based complex systems: lessons from ecology". *Science* 310, No. 5750, 987-991.
- Hamilton Environmental Services. 2013. *Winton Wetlands Index of Wetland Condition Assessments 2013-2013*. Benalla.
- Kok, K.; M. Patel; D. Rothman; and G. Quaranta. 2006- "Multi-scale narratives from IA perspective: part II". Participatory local scenario development. *Futures* 38, 285-311.
- Ladyman, J.; J. Lambert; and K. Wiesner. 2013. "What is a complex system?" *European Journal for Philosophy of Science* 3, No. 1, 33-67.
- Perez-Mujica, L.; T. Bossomaier; R. Duncan; A. Rawluk; M.C. Finlayson; and J. Howard. 2013. "Developing a sustainability assessment tool for socio-environmental systems: a case study of systems simulation and participatory modelling". *Proc. of the Int. Workshop on Simulation for Energy, Sustainable Development & Environment*.
- Perez-Mujica, L.; R. Duncan; and T. Bossomaier. 2014. "Using agent-based models to design social marketing campaign". *Australasian Marketing Journal (AMJ)* 22, No. 1. 36-44.
- Powell, S. J.; R.A. Letcher and B.F.W. Croke. 2008. "Modelling floodplain inundation for environmental flows: Gwydir wetlands, Australia". *Ecological Modelling* 211, 350-362.
- Scholl, H. J. 2001. "Agent-based and system dynamics modeling: A call for cross study and joint research". Paper presented at the System Sciences, 2001. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*.
- Sterman, J. 2001. "System dynamics modeling: Tools for learning in a complex world". *California Management Review* 4, 8-25.
- Tang, W., and D.A. Bennett. 2010. "The explicit representation of context in agent-based models of complex adaptive spatial systems". *Annals of the Association of American Geographers* 100, No. 5, 128-1155.
- Wilensky, U. 1999. Netlogo. Evanston, IL: Center for Connected Learning and Computer-Based Modeling. Retrieved from <http://ccl.northwestern.edu/netlogo/>
- Winton Wetlands Committee of Management Inc. 2011. *Winton Wetlands Restoration and Monitoring Strategic Plan*. Benalla: Retrieved from http://www.wintonwetlands.org.au/publications_news/publications/images/Restoration_and_Monitoring_Strategic_Plan.pdf.
- Young, O.; F. Berkhout; G. Gallopin; M. Janssen; E. Ostrom and S. Van der Leeuw. 2006. "The globalization of socio-ecological systems: An agenda for scientific research". *Global Environmental Change* 16 304-316

SIMULATING DAILY MOBILITY IN LUXEMBOURG USING MULTI-AGENT BASED SYSTEM

Hedi Ayed, Benjamin Gateau and Djamel Khadraoui
CRP Henri Tudor
29, avenue J.F. Kennedy, L-1855 Luxembourg
E-mail: firstname.name@tudor.lu

KEYWORDS

Daily Mobility Model, Methodology Simulation, Multi-Agent System, Urban Development

ABSTRACT

This paper describes a daily mobility model which allows realizing urban simulation to help understanding future land use management policies. The main challenge in such urban simulation consists of how to evaluate future situations since where comparison with real data cannot be possible. The work presented in this paper is about solving this problem and encouraging results have been obtained. Experiments have been made within different scenarios in Luxembourg.

INTRODUCTION

Luxembourg emerges as a very attractive cross-border regional metropolis leading to increasing residential migration and longer commutes. Empirical evidence shows that current urbanisation trends toward suburban and more remote per urban areas favour urban sprawl and car dependence. Urban sprawl is a very important territorial challenge to be addressed by policy makers since it is often associated with overconsumption of land and energy, fragmentation of natural habitats, difficulties in the provision of public services and increased residential segregation.

Further understanding social, economic and environmental impacts of both the residential mobility and the daily mobility of households is the core of this work. The MOEBIUS project relies on a solid knowledge base of interactions between daily and residential mobility acquired by the CEPS/INSTEAD, especially within the MOBILLUX project (Gerber, 2008). Within this work, we aim at developing an approach in order to simulate the future daily mobility (commuting patterns and travel mode choice) for different land use planning scenarios. When assessing those scenarios, particular attention is put on how they can provide a good trade-off between, on the one hand, economic growth via the provision of attractive and affordable living places, and, on the other hand, environmental, economic and social sustainability (modal split, land take, land rent, accessibility).

Using mathematical modelling and urban simulation can be applied on different domains such as transportation capacity, transportation system management, transportation demand management, land use/growth

management policies economic development policies and environmental policies. A good work for the introduction of urban simulation can be found in (Paul & Gudmundur, 2004). The origin of the application of computer simulation techniques for urban problems dates back to the 50s (E. Klosterman , 1994). Nowadays, the urban simulators are able to support the monitoring development as well as the objectives of the growth management acts. The MOEBIUS project is orientated towards a simulation of interactions between residential and daily mobility's for better understanding urban sprawl in Luxembourg and testing prospective planning scenarios. Given the complexity of the whole urban system (in particular the cross-border setting) and of these interactions, MOEBIUS is focused on the residential and daily mobility of the active population, i.e. workers employed in the Grand-Duchy and living inside the country.

METHODOLOGY AND APPROACH

Nowadays, simulations based on Multi-Agent Systems (MAS) are used in a growing number of sectors, where it is gradually replacing the various techniques of micro simulation and object-oriented simulation. This is due, in part, to its ability to capture very different styles of individuals, from very simple to more complex entities (as cognitive agents). The ease with which different levels of representation can be handled by the modeller is also one of its qualities comparing to the cellular automate. This apparent versatility makes MAS the medium of choice for the simulation of complex systems and spreads in an increasing number of domains: sociology, biology, physics, chemistry, ecology, economy, etc...

Simulation in Transportation Systems

Within the transport simulation, a list of extendible questions has to be taken into account: the vehicles movement, delays at crosswalks, time-dependence, and the itinerary that the agent will take, etc.

MATSim: a Multi-Agent Transport Simulator

MATSim is an open source multi-agent based transport simulation tool which was initially developed from the TRANSIMS project (Smith, Beckman, Anson, & Nagel, 1995) MATSim offers a framework for demand-modelling, agent-based mobility-simulation (traffic own simulation), re-planning, a controller to iteratively run simulations as well as methods to analyse the output

generated by the modules¹. MATSim represents each entity from the physical world (person, car ...) by an individual agent. The approach is mainly described by the following three concepts:

- The activity-day of an agent is described by a so-called plan. Iteratively, an agent tries to optimise his plan, by changing its intentions (leaving time, itinerary ...).
- The mobility simulation, which consists of the concurrently execution of the agent plans respecting limits set by the physics of the reality (speed limits, low capacity of roads, position of other roads, direction of roads, vehicle capacity, open time ...)
- Learning concept, which is responsible for making improvements of the agent choices. The system iterates between plans generation and mobility simulation. A function score is used to evaluate the performance of a plan.

Problem formulation

Specifically the problem we aim to solve, in this paper, is to design a platform around the MATSim simulator which will help to study and analyse the residential mobility across different scenarios. The objective is to measure the consequence of decisions, related to future developments, on citizen's mobility. This platform (with specific configurations is called "Daily Mobility Model" (see Figure 1).

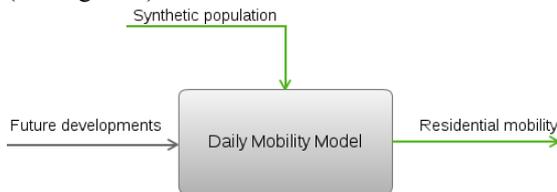


Figure 1: General Process

The daily mobility model is based on the modelling of route choice and mode choice for the journey to work. This includes (i) the agent specifications (i.e. age, type of household, car ownership...), (ii) the environmental characteristics (e.g. transportation networks, local densities at origin/home and destination/work), (iii) a decision tree (a set behavioural rules) obtained from a database describing mobility behaviours and used to simulate the mode choice. The behaviour of the agents will be controlled by MATSim, including interactions between agents like communication, and learning based on the previous route choices. The simulation of route choice will consist of executing the "best path".

By scenario we mean, firstly, a strategic tool for choosing between different possibilities of development and, secondly, a tool for the coordination of sectorial plans as well as a framework for planning at the regional and communal levels. The scenarios will primarily vary the location of expandable land (according to accessibility measures) and threshold densities (within a raster GIS). In addition we may vary

workplace clusters considering specific development projects.

A synthetic population may be defined as an artificial population composed of individuals with associated individual characteristics (level n), and constructed from known at the aggregate level (n+1) census data. This population is called "synthetic" because it is reconstructed using conditional probabilities and assignment algorithms. It considers each individual person as an "agent" by exceeding the aggregate data provided by the statistics. These agents, whose characteristics are deduced from the general census of the population, are grouped into households, these households, whose number and size are aligned on demographics, "retrieve" additional features that emerge from individual agents (this is particularly the case for household income, calculated from the occupational category of individuals who compose it). In this work, the synthetic population of Luxembourg has been constructed based on demographic perspectives (time horizon 2020) and available households' surveys at a spatially disaggregated level. Barthelemy and Cornelisi (Barthelemy & Cornelis, 2012) explain the most known synthetic population generation methods.

THE DAILY MOBILITY MODEL

The approach that we propose in this paper, is defined on two steps. The objective of the first step is to validate the configuration of the MATSim simulator with real data, this step is called the t_0 (stage of validation). By t_0 we mean the current situation regarding infrastructure and residential place of the synthetic population. MATSim will require several configuration parameters to get accurate results, for this reason we consider the t_0 as a capital phase. Indeed, this step consists of configuring the simulator, and then validates these configurations with a simulation using real data. The validation of this simulation is possible through real data based on the results of the cross-border transport survey held in 2003 by CEPS/INSTEAD. The configuration step will be repeated until the validation conditions are meet.

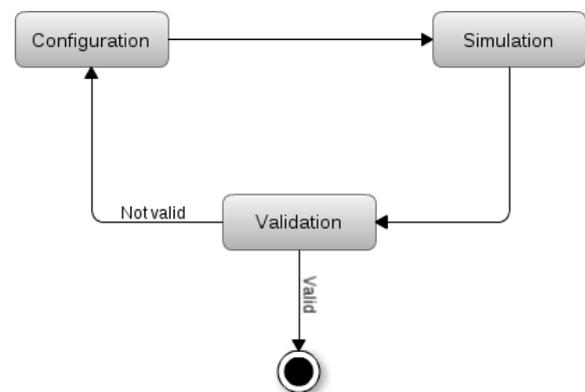


Figure 2: The processing steps of the t_0 stage

The second step consists of making the main simulations of the different scenarios. Since the configuration steps is performed, we assume that the

¹ <http://www.matsim.org/>

model is now able to provide realistic simulation based on the input data of the simulated scenarios and the configuration parameters resulting from the first step.

We call this second step t_1 (stage of execution). In MATSim, the result of a simulation is a set of events, which are used to document changes in the state of an object (agents, vehicle, arc, edge ...). This list of events is then used in the post-processing step in order to analyse the simulation and create the output of the daily mobility model. The output put of the model will be discussed below.

The validation stage: t_0

Every simulation is based on a theory, namely a set of postulates, laws, principles, theorems ... to build a model from data and initial assumptions then used to produce the outcome. In our case, the theory (traffic rules and law, planning, etc. ...) is provided by MATSim and data are the synthetic population and the infrastructure of t_0 . The main configurations of the MATSim simulator are carried out through the configuration input file. Following are values of the parameters that we have configured within this stage.

MobSim

MobSim is the execution module of the agents' plans, several MobSims have been developed, some of them are still in use, and others are obsolete². The MobSim we have used in this work is the QSim, which is a queue model time step approach using a deterministic algorithm³. The QSim is an extended version of the QueueSimulation. Features like traffic signals, within-day and public transport are available with the QSim implementation.

Replanning Rates

The replanning is an important step on the MATSim iterative processing; it consists of changing the behaviour of the agent in order to optimize his day plan execution. This change can affect his itinerary by choosing a new route or he structures his current plan (e.g. the start and end-time of an activity), sometime the agent keep the same plan for the next iteration. A replanning module is a component which applies a specific strategy to adapt the agent plan. Within MATSim it's possible to develop (modify) a personal module and some default replanning modules are available. Different modules can be used within the same simulation, in such situation the agent will choose one module randomly based on a configuration parameter.

The figure 5 presents the configurations of the strategy module that we chose. Each probability module is defined by two parameters. The first one specifies the

name of the module. And the second one defines the probability that the strategy will be adopted by the agent. After a manual optimization step we have set the configuration values as following:

- BestScore: This module will select the plan which has the best score from the existing plans of the agent. This module has 80% of chance to be selected
- ReRoute: Try to find a new optimal route by calculating a new one using the default routing algorithm, travel times are selected from the previous iteration. This module has 10% chance of being selected
- TimeAllocationMutator: Changes randomly times of the agent's plan, by changing the duration of an activity within the respect of the person specification. This module has 10% chance of being selected.

```
<module name="strategy">
  <param name="maxAgentPlanMemorySize" value="5" />
  <param name="ModuleProbability_1" value="0.8" />
  <param name="Module_1" value="BestScore" />
  <param name="ModuleProbability_2" value="0.1" />
  <param name="Module_2" value="ReRoute" />
  <param name="ModuleProbability_3" value="0.1" />
  <param name="Module_3" value="TimeAllocationMutator" />
</module>
```

Figure 3: The replanning strategy con

The objective of the replanning phase is to optimize the day plan of each agent and so the system became more realistic and react accurately the behaviour of the population we want to simulate. This is due to the fact that naturally the human always try to optimize his itinerary and finally chooses the best path according to his preferences (arrival time, modes ...). The diversification and the intensification are two important aspects in the field of the optimization. Diversification means the process of gathering information about the problem; this can help to discover more possible solutions. Intensification (or exploration) aims to use the information already collected to define and browse the interesting areas of the search space. The concepts of intensification and diversification are preponderant in the design of combinatorial algorithmic, which must achieve a delicate balance between these two dynamic searches. In our case, this may be controlled using the strategy module configuration. Indeed the "best score" module aims to choose the best plan according to the score function, and so it promotes the intensification. The "ReRoute" and the "TimeAllocationMutator" will try to discover new possible solutions (best day plan) by doing some minor modifications within the current plan, this can promotes the diversification.

Change Scoring Parameters

The score of a day plan is the objective function in optimization. "Objective function" is the term used to describe a function that serves as a criterion to determine the best solution to an optimization problem. Specifically, it associates a value to an instance (a solution) to an optimization problem. The goal of the

² more details can be found in: An Overview of the MobSim; <http://matsim.org/node/619>

³ A deterministic algorithm is an algorithm that at each step always go to the next step in the same way

optimization problem is then to minimize or maximize this function until the optimum.

The principle of the scoring function of MATSim is described in detail in (Charypar & Nagel., 2005). The Figure 4 describes the parameters that we used to configure MATSim scoring function for the t_0 stage. The signification of those parameters is the following:

- **lateArrival** : Decrease the score by 18 units if the agent arrives late to his last activity (work in our case)
- **earlyDeparture**: No influence on the score if the agent starts his route early. It can be logic if it's the only way to respect his arrival time
- **performing**: Increase the score by 6 units if the agent arrives to the last activity with respecting all the time constraints. In our case the departure time from home and the arrival time to work.
- **traveling**: Decrease the score time of plan when the agent is traveling. This encourages the agent to arrive as soon as possible.
- **waiting**: No problem if the agent needs to wait.

```
<module name="planCalcScore">
  <param name="learningRate" value="1.0" />
  <param name="BrainExpBeta" value="2.0" />

  <param name="lateArrival" value="-18" />
  <param name="earlyDeparture" value="-0" />
  <param name="performing" value="+6" />
  <param name="traveling" value="-6" />
  <param name="waiting" value="-0" />
</module>
```

Figure 4: The scoring parameter configuration

The routing algorithm

By default, MATSim uses a routing algorithm based on Dijkstra's shortest path algorithm; however, it is also possible to use another type of algorithm. An implementation of the A* algorithm is also available with MATSim, which is a heuristic extension of the Dijkstra' algorithm known to be faster. We used the A* algorithm as a routing algorithm, this can be done with the "routingAlgorithmType" parameter of the "controller" module, see Figure 5.

```
<module name="controller">
  <param name="outputDirectory" value="6OUTBASE;" />
  <param name="firstIteration" value="0" />
  <param name="lastIteration" value="100" />
  <param name="runId" value="run0" />
  <param name="routingAlgorithmType" value="AStarLandmarks" />
  <param name="mobsim" value="qsim" />
  <param name="writePlansInterval" value="1" />
</module>
```

Figure 5: The "controller" module configuration

The to input network

Within MATSim, one of the possible sources of the transport network is the OpenStreetMap⁴; indeed MATSim offers some tools to transform an OSM (OpenStreetMap) map to a MATSim network. The OSM map that we used for this stage, came from the Geofabrik's free download server⁵, this server extracts

data from the OpenStreetMap project which are updated every day.

The to population

The synthetic population of t_0 represents the current real population, with a day home to work scenario of the Luxembourg resident. The decision to exclude the border is due to the unavailability of data and not for technical reasons. The population is initially composed by about 190.000 persons, after applying the modal choice using the decision tree, 76.9% of them choose to use the car mode. The input plans file is then composed by about 146.000 agents using their car. The time planning of activities (home, work) is as follows:

- **Leaving home**: The time, when the agent should live his home is randomly selected between 06h00 and 08h00.
- **Starting work**: An agent is expected at work between 06h00 and 10h00. This time is selected randomly according to a fixed probabilities (see experiments section)

Those values were chosen based on the result of Luxembourg transport survey held in 2003. We should also note that the plan generator takes into account the distance between the home and the work to generate the start and the arrival time, the expected itinerary should be feasible in this range of time. Despite this, some agents will not be able to respect their plans due to the congestion and then have to find an alternative (an early departure or a late arrival for instance). This choice will depend of the plan score.

The execution stage: t_1

The main simulations of the different mobility situations will be performed within this stage; this will be done by developing scenarios using official strategic planning documents to represent the promoted and the real-world visions of regional planning. For more details on the differences scenarios please refer to (Rieser, 2010).

The method for developing scenarios that articulates different sets of available lands is carried out with the use of data that can be found in usual municipal/agglomeration governments. Its implementation is based on a Geographical Information System (GIS) where planning scenarios are built on the "interpretation" of the Luxembourg official urban planning documents combining different scales: 1) national development framework, 2) regional planning orientations, 3) sectorial development plan, and 4) local management plan for residential development.

Following the scientific literature and the review of spatial planning policies within the context of Luxembourg we identified four spatial development scenarios: 1) Inner City Development, 2) Transit Optimization Development (TOD) 3) Centers of Development and Attraction System (CDA) and, 4) Business as Usual (BaU). These scenarios are based on the dimensions derived from the scientific literature on

⁴ <http://www.openstreetmap.org>

⁵ <http://download.geofabrik.de/>

sustainable urban development and the specific objectives of the spatial planning visions.

The modal choice

The module of the transport mode prediction is based on the work of Omrani et al. (Omran, Charif, Gerber, & Awasthi, 2013). The objective of this module is to select a mode for an agent based on the characteristics of the person that it models. This module is based on an Evidential Neural Network (ENN). The presented model uses individuals' characteristics, information on the daily mobility, transport mode specifications and data related to places of work and residence. The results were compared by cross-validation the rates of successful prediction obtained by ENN and several alternative approaches. The results show that the ENN is superior to the studied alternatives. The outcomes of this module are a set of behaviour rules. Those rules help to build the decision tree (see an example in Figure 6, where nodes are probabilities of each mode to be selected (car, PT: public transport and others: bike, walk ...) and branches present the value of the tested attribute (distance from home to work, gender, etc.). The decision tree is then used by the plan generator to select randomly a mode to the person according to his characteristics. The generator will determinate the node in which the person belongs and so the probabilities of each mode. Finally the returned mode will be selected according to this probability.

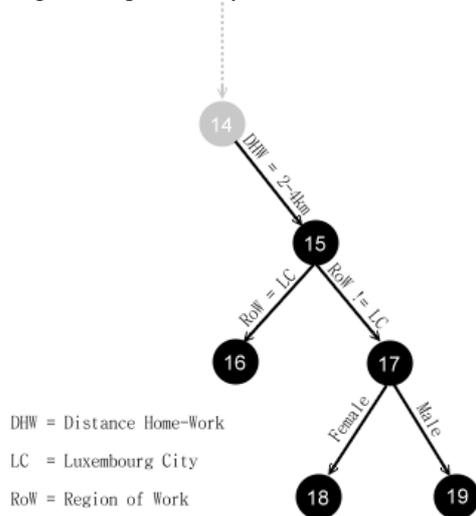


Figure 6: A part of the modal decision tree. (Probabilities is as following: Node 16: 73% car, 10% PT, 17% others. Node 18: 70% car, 13% PT, 17% others Node 19: 72% car, 11% PT, 17% others)

Future developments

In order to simulate the future daily mobility with different land use scenarios, we should take into account the future urban developments. Otherwise it's not accurate to test future population mobility within a current transport network. In order to consider the future planned routes we used an OpenStreetMap map which allows taking under consideration roads that are in constructions or planned in near future. This feature is available by using the "proposed" and "construction"

tags. Some other tags can also be used to detail this feature, namely the date of the expected end of constructions. So, to build the future network we have used the current OpenStreetMap map of Luxembourg by changing all roads that are under construction (or planned) to be considered as in-use. Only constructions that are expected before 2020 have been integrated. The obtained network (t_1) is composed by more than 450 new links and 200 nodes.

EVALUATION OF SIMULATION RESULTS

The daily mobility model proposed in this work is organized in two phases. The aim of the first is to configure the simulator and to validate these configurations within a real data. This first stage is very consequential on the reliability of the simulation results within the second stage. In the previous section we have presented the configurations approved in the first stage. In the next paragraph we will present the simulation results in terms of traffic analyses and comparison with the real situation. Since the objective of this stage is only the configurations of the model, no discussion on the simulation performance will be presented. This work will only focus on the agents' behaviours to compare with the real case.

Results of stage t_0

The simulation was performed until meeting the optimum, which means stability on the variation of the score agents (after 30 iterations). The simulator output was then analysed to deduct, for each agent, the executed route plan in the last iteration (the optimal) and therefore its itinerary, departure time, arrival time and the transport mode (necessarily the car in our case). Next, these results were crossed with geographic data to determine the time spent by agents to travel from a given residential commune to work place. Table 1 presents a comparison between times obtained by the model and real times resulting from surveys. The table presents the mean, median, minimum and maximum time to travel from a home region to the different work regions. The chart presented by the Figure 7 resumes the results of the previous table by giving the error rates of obtained times relative to real times.

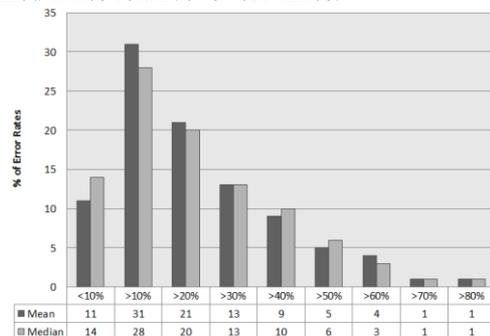


Figure 7: Error rate of mean and median simulated times relative to real times for t_0 .

Table 1 The results of the t_0 Simulation: comparing the agent travel time with real times

Regions		MATSim Simulation Results				Survey Times			
Home	Work	Mean	Median	Min	Max	Mean	Median	Min	Max
North-Center	Basse Alzette	8,21	8.0	0	30	11.46	10	2	30
North-Center	Center	28,14	28	3	60	29.94	30	10	90
North-Center	Moselle	27.8	28	5	60	29.7	26	15	60
North-Center	Oesling	21.79	21	2	57	21.54	20	10	35
North-Center	South	42.06	42	20	191	48.57	40	40	90
North-Center	Luxembourg City	33.15	33	12	61	34.02	35	1	60

Results of stage t1

First the table 2 presents the modal distribution according to the different used scenarios. As the population is quite different from one scenario to another, the modal distribution is neither the same, as it can be seen in table 2. This modal distribution directly affects the number of agents that will take part to the simulation (only persons taking their car are represented). Those scenarios have been simulated separately by running the same simulator environment (configurations outcome from the first stage) within specific agents (representing the input scenario) at each run, as described in Figure 3. So, for simplicity and without loss of generality, to analyse the MATSim output simulation results, we will take the "Business as Usual" scenario as an example.

Table 2: Modal distribution of the different scenarios

Scenario	Car(%)	PT(%)	Others(%)	Agents(car)
Inner City	71.67	18.55	9.77	150849
TOD	73.76	17.97	8.25	157705
CDA/Luxembourg	76.13	17.54	6.31	159934
CDA/Belval	76.32	16.76	6.90	162354
BaU	77.36	17.06	5.56	162650

One of the important analyse offered by the MATSim controller, is the score statistics generator, which may help to follow the progress of the optimization. Figure 8 presents a chart of best, worst, average and executed plans by agents during a complete simulation (30 iterations). It is especially important to note, that the curve scores of the executed plans is stagnating within the 20-25 iterations. This finding is very important in the optimization field, because it confirms the convergence of the model. It is of course very important to ensure the convergence of an algorithm, but the speed of convergence and complexity are also factors to consider when designing or using a model. Within MATSim, the configuration step may have a strong influence on the model convergence.

Now let's focus on the behaviour of the agents in term of expected departure and arrival times. As we have explained in the configuration section, the MATSim controller will try to change the departure time of the agent if there is no way to optimize the score function, in other terms if there are many planned congestions on the taken road. We have allowed this behaviour by configuring the early departure to have no influence on the score function. However, this can conclude to

inaccurate results if the number of agents that does not respect the departure time is important. First, the Table 3 gives the number (and percentage) of agents leaving the home before 06h00 and those after 08h00. This time slot has been fixed in the configuration step. More than 90% of agents respected the departure time condition. The table presents also a comparison between the percentage of expected agents by arrival time slot (this percentage is fixed by the configurations) and times obtained by the simulation.

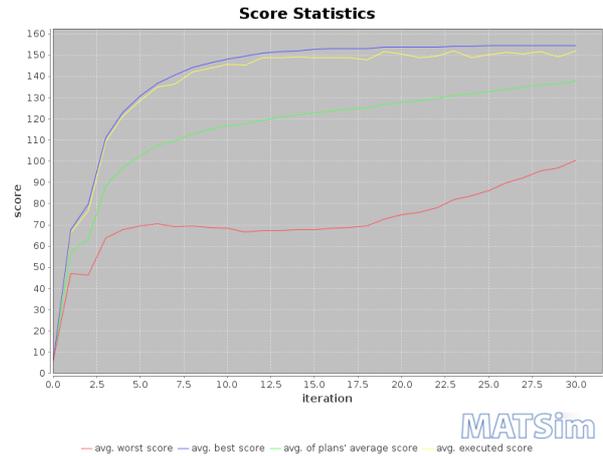


Figure 8: Statistics of the score function for the "Business as Usual" scenario simulation

CONCLUSIONS

This paper has sought to develop an operational urban simulation model called "a daily mobility model", in order to further understand the social, economic and environmental impacts of the residential mobility and the daily mobility of households, with a specific reference to the MATSim simulator. To be relevant, our challenge was the design of a model that should be able to work with future scenarios and gives reliable results. To deal with this problem, we have presented a model composed by two steps. The first one consists of a configuration process in which we have iteratively simulated the behaviour of the current population and compared the output of the model with the real data until reaching a threshold of acceptable similarity at which the model is considered reliable.

Table 3: Departures and arrivals time intervals: Expected and Simulated

	Departure Time		Arrival Time			
	Before 06:00	After 80:00	Before 06:00	06h – 08h	08h – 09h	09h – 10h
Number	3035	12496	331	116490	43188	2634
Percentage (simulated)	1.87%	7.68%	0.20%	71.62%	26.55%	1.62%
Percentage (expected)			0.00%	70%	25%	5%

The second stage defines the process of the core simulation to perform, in this paper we have presented the simulation results of several land use planning scenarios of Luxembourg. We have also presented in this paper the transport modal choice used by the agent to select their principal transport mode as well as their preferences implemented as a score function in MATSim.

The daily mobility model will be further developed to add the transport modal choice into the simulator, this will help to get a more realistic modal distribution by letting the agent choose their mode according to their benefit. Also taking into account the public transport mode is part of our upcoming perspective, this includes especially the park and ride modality which is very important within the Luxembourg context.

REFERENCES

- E. Klosterman , R. (1994). Large-Scale Urban Models Retrospect and Prospect. Journal of the American Planning Association.
- Barthelemy, J., & Cornelis, E. (2012). Synthetic populations: review of the different approaches. *CEPS/INSTEAD*.
- Charypar , D., & Nagel., K. (2005). Generating complete all-day activity plans with genetic algorithms. *Transportation*, 369-397.
- Gerber, P. (2008). *Comprendre les interactions entre les mobilités quotidienne et résidentielle au Luxembourg et son aire métropolitaine transfrontalière*. Luxembourg: Rapport scientifique final pour le FNR, 120 p, Luxembourg.
- Omran, H., Charif, O., Gerber, P., & Awasthi, A. (2013). *Prediction of individual travel mode using evidential neural network model*. CEPS/INSTEAD.
- Paul, W., & Gudmundur, F. (2004). *Introduction to urban simulation: Design and development of operational models*.
- Rieser, M. (2010). Adding Transit to an Agent-Based Transportation Simulation: Concepts and Implementation. *PhD thesis*.
- Smith, L., Beckman, R., Anson, D., & Nagel, K. (1995). TRANSIMS: TRAnspOrtation ANalysis and SIMulation System . *National transportation planning methods applications conference*.

AUTHOR BIOGRAPHIES

- Dr. Hedi Ayed has more than 5 years' experience in algorithm design and optimization. Following a Ph.D. in computer science he took a position as computer scientist at CRP Henri Tudor. He works in the areas of transport, intelligent systems and software engineering. His recent interests are related to traffic simulation and multi-agent systems. He is expert in the application of the optimization algorithms to problem is related to mobility. He mainly participated in several previous and current EU projects: CARLINK, MOEBIUS, STIMULATE, ELECTRA.
- Dr. Djamel Khadraoui received his PhD in Vision for Robotics (1996) from University Blaise Pascal (France). He is at CRP Henri Tudor since 2002 as a Program Manager of Mobility activities. He is Lead R&D Manager and active in the areas of eMobility and Critical Infrastructures Protection. His main scientific interests are: intelligent and adaptive systems, distributed systems (multi-agents systems), software engineering, trust and security in IT Infrastructures. He mainly participated and coordinated several national and EU projects. He is currently providing lectures at the master level in IT Security (Distributed Systems Security, PKI, DB security and Dependability) at the University of Luxembourg and University of Lorraine (France). He co-supervised several PhD works related to multi-agents and distributed systems, security, optimization techniques for telecommunication and multimodal transport issues.
- Benjamin Gateau received his Ph.D. in Computer Science at Ecole Nationale Supérieure des Mines de Saint-Etienne in June 2007 resulting from a joint project between the CRP Henri Tudor -- Luxembourg and the Ecole Nationale Supérieure des Mines de Saint-Etienne -- France. He currently R&D Engineer at CRP Henri Tudor. His research interests are Distributed Systems and mainly Electronic Institution Model for Multi-Agent Systems, Supervision of SLA for Cloud Computing and IoT in the context of smart buildings.

AGENT-BASED CONTROL FRAMEWORK IN JADE

Franco Cicirelli, Libero Nigro, Francesco Pupo
Laboratorio di Ingegneria del Software
Dipartimento di Elettronica Informatica e Sistemistica
Università della Calabria
87036 Rende (CS) – Italy
Email: f.cicirelli@deis.unical.it, {l.nigro,f.pupo}@unical.it

KEYWORDS

Multi-agent distributed systems, control extensions, concurrent/parallel execution, real-time, simulation, JADE, FIPA, Java.

ABSTRACT

This paper proposes an agent-based control framework in JADE which allows the construction of control extensions tailored to the application needs. The approach is based on a minimal actor model which simplifies JADE agent programming. A catalog of reusable control forms, both concurrent/parallel and time-dependent (real time or simulation time are supported) was achieved. The paper introduces the control framework, clarifies its implementation status and demonstrates its practical use by examples.

INTRODUCTION

This work develops an agent-based control framework in JADE (Java based Agent DEvelopment framework) (Bellifemine *et al.*, 2007)(Jade, on-line) whose aim is to enable both experiments of mechanism design (Wooldridge, 2009) and the definition of application specific control structures, e.g. based on discrete-event simulation. The proposal tries to widen the applicability of JADE to time dependent and possibly high performance applications.

JADE was chosen because it is based on Java, it is an open source middleware for distributed computing, it adheres to FIPA (Foundations for Intelligent Physical Agents) specifications (Bellifemine *et al.*, 2007)(FIPA, on-line) which in turn open to interoperability with compliant general-purpose legacy software (e.g. a visualization service useful in a simulation application). In addition it supports agent mobility and permits an exploitation of the computing potential of modern multi-core machines.

The control framework relies on a minimal actor model of computation (Cicirelli *et al.* 2009, 2013a) and on the concept of a control structure which has a reflective link and controls the evolution of a collection of cooperating actors. The actor model simplifies the implementation of JADE agents by hiding the underlying behaviour framework. The control structure transparently filters message exchanges among actors and superimpose to them a suitable delivery policy which ultimately depends on the application goals.

Multiple control structures possibly running on different computing nodes (JADE containers) can coordinate to one another so as to ensure time synchronization in a distributed simulation scenario (Cicirelli *et al.*, 2009, 2014) or the fulfillment of event precedence constraints due to causality consistency or causal delivery (Tanenbaum & Steen, 2007) in general distributed systems.

The work described in this paper differs from known JADE based simulation tools (e.g. (Derksen *et al.*, 2011)(Gianni *et al.*, 2008)) for the openness and flexibility of the proposed approach. Discrete event simulation is only one kind of an achievable control form. A major benefit of the JADE proposed approach consists in the possibility of configuring, e.g., an agent-based simulation and let it to run in a container launched on a high-performance remote machine or in the cloud. All of this stems from the distributed character of the JADE middleware.

This paper introduces the control framework, describes its implementation status, and demonstrates its practical usefulness and programming style through modelling examples. Lessons learned from the implementation experience are also highlighted. Remarks about on-going and future work are discussed in the concluding section.

BASICS OF AN ACTOR MODEL

A variant of the Actors model (Agha, 1986) is considered in this work. Actors hide internal data variables and have a behavior (finite state machine) for responding to messages. The communication model consists of asynchronous message passing. An actor is a reactive entity which answers to an incoming message on the basis of its current state and received message. An actor is at rest until a message arrives. Message processing is atomic and triggers a data/state transition. Basic operations of actors include:

- *new*, for creating a new actor
- *become*, for changing the actor state
- (non blocking) *send* for transmitting messages to *acquaintance* actors (including itself for proactive behavior). The send operation carries a message with a timestamp which specifies when the message has to be delivered to its recipient.

The evolution of a subsystem of actors running on a single processor is regulated by a control machine, which is in charge of (transparently) scheduling (buffering) sent messages and applying to them an application tailored

(e.g. time sensitive) control discipline. The control machine repeats a basic loop: at each iteration one message is selected in the set of pending messages, and delivered to its target actor by causing its handler(msg) method to be executed with the message passed to it as an argument.

Theatre (Cicirelli *et al.*, 2009, 2011, 2013b, 2014) is a distributed architecture based on actors. Each theatre hosts a collection of local actors and a governing control machine. Theatres of a distributed actor system can coordinate to one another so as to ensure the fulfillment of event causality consistency or time synchronization in a distributed simulation scenario. Theatre was successfully implemented on top of different transport layers such as Java Socket (Cicirelli & Nigro, 2013b), HLA (Cicirelli *et al.*, 2009, 2011, 2014), etc.

JADE CONCEPTS

JADE is an open source Java-based framework for the development of distributed multi-agent systems. It acts as a middleware which hides heterogeneity in a distributed context, and provides the runtime support to agents.

Agents execute in the so-called containers organized into platforms. A platform (distributed system) is booted by starting a main-container which hosts some fundamental agents providing services (for naming, mobility, information sharing through “yellow-pages” etc.) to user-defined agents. Other containers can then be launched to join with an existing main-container thus establishing a given platform.

The launch of a container can be accompanied by the start of some agents. Agents can also be created through the RMA (Remote Management Agent) GUI, platform specific. Finally, agents can dynamically be created as part of the application logic.

JADE agents are thread-based. They receive messages via a local mailbox (Agha, 1986) and process them one at a time through a behaviour structure. Ready to use base behaviours can easily be adapted to the modelling needs. Complex behaviours (e.g. sequential or parallel) are also available. Behaviours can be added to an agent dynamically.

The JADE communication model is based on asynchronous messages expressed using FIPA ACL (Agent Communication Language). A message can carry either simple textual information, or complex serialized Java objects. An application can rely on a family of ACL messages sharing a certain ontology, that is a domain specific vocabulary.

JADE agent programming is supported by some Java classes/interfaces like Agent, Location, AID (Agent unique Identifier), Behaviour, ACLMessage along with associated attributes and methods.

It is worthy of note, though, that no primitive support exists in the API for a notion of time or of mechanisms for building a simulation model. The threaded agents, on the other hand, are not ideal for implementing a simulation infrastructure where the controlling entity, which is responsible of time management, cannot proceed

with the next decision about the event/message to fire without knowing the current agent has finished processing its received message. All of this motivated the work described in this paper aimed at making it possible to experiment with general control extensions in JADE.

A CONTROL FRAMEWORK IN JADE

The adopted actor model can be minimally supported by a lightweight architecture where actors are thread-less objects and one thread of control is held by the control machine of a given theatre. This thread supports execution of the message handlers of the actors. Other threads can be present in a theatre for input/output message communications with partner theatres.

In this work a theatre is mapped onto a container, although nothing forbids the use of multiple containers to host the actors assigned to the theatre. In particular the focus will be on a single theatre model. Multi-theatre applications in JADE are beyond the scope of this paper.

Actors and control structures are uniformly based on JADE agents. Messages are extensions of FIPA based ACLMessage base class. As a consequence, interactions between actors and a control machine are (transparently) realized through the exchange of suitable ACLMessage(s).

The following class hierarchy was developed. Base (abstract) class Actor inherits from jade.core.Agent and exports the following methods:

- *void become(int status)* – changes the actor internal (control) state
- *void send(Message m)* – sends m to its recipient
- *int currentStatus()* – returns the actor current state
- *void handler(Message m)* – processes m
- *void bind(String control)* – links this actor to the control structure whose name is control
- *void setup()* – installs the actor behavior
- *void newActor(String nick, String className, Object[] args)* – creates dynamically a new actor
- *void startControl()* – launches the control this actor is bound to. Used by a master/configurator actor only.

The Actor class provides a (hidden) cyclic behavior to heirs. In the action() method of this behavior a message is received from the actor mailbox and the handler of the receiver actor is then launched. At handler termination, the action sends a message to the control machine informing about the end of message processing.

The resultant programming style of actors in JADE is simple yet effective as will be shown later in this paper. Purposely, actor programming hides the internal behavior details.

The base class Message derives from ACLMessage. It carries a timestamp useful for time-dependent control forms, and a validity flag which affects actual message dispatching. At its creation, a message gets the receiver AID. Getter/setter methods are available for managing the receiver, the timestamp, and the validity flag attributes.

The (abstract) class `ControlMachine` derives from `core.jade.Agent`. A particular control machine derives from `ControlMachine` and implements a certain control form through its behavior structure. A library of control structures was developed which includes `Concurrent`, `Parallel`, `Simulation` and `Realtime` forms. Other control mechanisms can be added as well.

`Concurrent` implements co-operative concurrency based on actor-handler interleaving. Handler execution is atomic. Messages are processed one at a time thus easily fulfilling any precedence constraints. The control structure is based on a FIFO message queue (MQ). `Concurrent` takes down when application messages exhaust or a specific termination message (with `CANCEL` performative) is received.

`Parallel` control form enables parallel execution of message handlers and can improve execution performance with respect to `Concurrent` by allowing an exploitation of a multi-core architecture and of its hyper-threading feature. `Parallel` assumes that actor handlers have no precedence constraints and thus can be executed in any order. `Parallel` operation roughly corresponds to that of a thread-pool, where tasks are submitted by dispatching messages to actors. Conceptually, no message buffering is accomplished in the `Parallel` control structure. In addition, no implicit termination condition is recognized.

Experimenting with the above control examples has highlighted a critical problem in JADE concerning the creation of dynamic actors. The `Actor.newActor()` method gets the container controller and then invokes on it the `createNewAgent()` method. In the operation of a `Concurrent`-based application, there is no problem with these operations. However, in a more congested situation like that of a `Parallel`-based scenario, the above-mentioned basic operations must be synchronized, thus limiting the achievable execution performance.

`Simulation` follows the same interleaved handler() execution of `Concurrent`. In addition a virtual time notion of a classical discrete-event simulation schema is maintained. Messages with absolute timestamps are buffered into a `java.util.PriorityQueue` collection (time queue or TQ), where the head message holds the (or one with) minimum timestamp. At each iteration of the control loop (action method of the control behavior), the most imminent (in time) message is extracted from TQ and dispatched to its receiver. Then the behavior expects an `INFORM` message communicating the handler termination. The behavior of `Simulation` terminates when either TQ empties or the virtual time exceeds the assumed simulation time limit. The use of `Simulation` is assisted by a package (`actor.distributions`) of common density distribution functions (including uniform, exponential, hyper exponential, erlang, normal etc.) based on `java.util.Random` pseudo-random number generators.

`Realtime` is another time-dependent control form useful for real time applications. It rests on a real time notion achieved on top of `Java System.currentTimeMillis()`. Messages can be or not time-constrained. Not time-

constrained messages (created without an explicit timestamp) are assumed to be processed in FIFO order and when there are no fired time-constrained message. Time-constrained messages must be dispatched as soon as the current time exceeds their firing time. Timestamps are specified by relative times with respect to current (implicit) time. Absolute fire time is generated by the control form. Two message buffers are used: MQ as in `Concurrent`, and (a priority queue) TQ similar to `Simulation`. `Realtime` control loop is never-ending. In the case there are no messages in MQ and current time is lesser than the most imminent message in TQ, the control structure simply awaits current time to advance to the firetime of the first message in TQ. Obviously, the behavior of `Concurrent` is available as a particular case of operation of `Realtime`.

The use of `Realtime` implies some interface actors to the external environment (perceived by sensors and operated by actuators) are to be introduced so as to transform environment stimuli into internal messages and vice versa.

PROGRAMMING STYLE

To figure out the resultant programming style of actors in JADE, the following shows a modelling example based on concurrent/parallel control. The example is concerned with an agent-based version of classical “divide et impera” merge sort algorithm which rests on a binary tree of recursive method activations (frames). Recursion here is replaced by agent creation and message passing. Each agent sorts a distinct subvector and sends to its parent a done message when it has finished. The parent in turn, when both the two done messages from its left and right childs are received, merges the two sorted subvectors and sends itself a done to its parent, if there are any, and so forth. The application is designed for a shared memory context, where all the sort agents execute on a same theatre/container launched on a standalone (possibly multi-core) machine.

Bootstrapper Actor

A mergesort agent (MSA) is used for the booting process (see Fig. 1). It can be created from the RMA GUI of a JADE main-container launched by a command line along with the control agent (e.g. `Concurrent`) thus:

```
>java jade.Boot -gui cc:actor.Concurrent
```

Of course, other solutions are possible such as choosing the control structure from the yellow pages service of the platform. The MSA agent introduces and fills the array to be sorted, creates the first (root) Sorter agent with the control form as an argument, and creates and sends to it an activation `Sort` message (see Fig. 2) which carries the receiver AID, null as the parent AID, and the specific subvector to sort.

The algorithm of MSA is coded in its `setup()` method (required by JADE). The first action of `setup()` consists in

invoking the same method of the super class (i.e. Actor). This is mandatory for properly installing the behavior structure of the actor. MSA is then bound to the control form through the bind() method. The last action of setup() starts the control agent. Since MSA, after booting the application, no longer takes an active role in it, its handler() method was redefined with an empty body.

```
package actor.mergesort;
import jade.core.AID;
import actor.*;
public class MSA extends Actor{
    private int a[]=new int[200000]; //example
    public void handler( Message m ){
    public void setup(){
        super.setup();
        Object[] args=getArguments();
        if( args==null || args.length==0 || args.length>1 ){
            throw new RuntimeException("MSA wrong arguments.");
        }
        //get name of control structure
        String control=(String)args[0];
        bind( control ); //binds this actor to its control form
        Object[] args={control}; //prepare arguments for root
        for( int i=0; i<a.length; ++i ) a[i]=a.length-i;
        System.out.println( java.util.Arrays.toString(a) );
        //create root Sorter
        newActor("root","actor.mergesort.Sorter",args);
        AID raid=new AID( "root", AID.ISLOCALNAME );
        Sorter.Sort s=new Sorter.Sort(raid, null,a,0,a.length-1);
        send( s ); //send activation message to root
        startControl();
    } //setup
} //MSA
```

Figure 1 – The bootstrapper MSA actor

For the sake of simplicity, the Sorter actor class in Fig. 2 is reported partially in pseudo-code. The actor admits two message classes: Sort and Done, the latter being restricted to an internal use only. A user-defined message class must always be provided of the receiver AID which is immediately passed to its super class Message. The actor behavior (see the handler() method) can find itself into the SORTING or MERGING state. In the initial SORTING state the agent gets initialized by receiving a Sort message.

For efficiency, in the case the subvector $v[inf..sup]$ has no more than, e.g., 50 elements, it is immediately sorted by selection sort. Larger sizes are handled “recursively” by creating left and right child actors and initializing them with corresponding subvectors. Note that the current agent is established as parent of child agents. After divide-et-impera, the Sorter passes to the MERGING state where two Done messages are expected. On receiving the done messages from childs, the actor does the merging task, after which it communicates it has finished by sending a Done message to its parent. In the case the parent is null (root actor), the array is printed. It should be noted that although actors normally have a non terminating behavior, in this case, after the sort process is finished (e.g. after a selection sort or a merging phase) the agent executes doDelete() to self-destroy and memory reclamation.

As a final remark, the necessity of using unique names for dynamically created actors has to be pointed out. In Fig. 2 child names are built by taking into account the values of inf, med and sup indexes.

```
package actor.mergesort;
import jade.core.AID;
import actor.*;
public class Sorter extends Actor{
    private int[] v;
    private int inf, med, sup, doneCount;
    private AID parent;
    private String control;
    public static class Sort extends Message{
        int inf, sup;
        AID parent;
        int[] v;
    public Sort( AID receiver, AID parent, int[] v, int inf, int sup ){
        super( receiver );
        this.parent=parent; this.v=v; this.inf=inf; this.sup=sup;
    }
} //Sort
private static class Done extends Message{
    public Done( AID receiver ){ super( receiver ); }
} //Done
private static final int SORTING=0, MERGING=1; //states
public void handler( Message m ){
    switch( currentStatus() ){
    case SORTING:
        if( m instanceof Sort ){
            parent=((Sort)m).parent; v=((Sort)m).v; inf=((Sort)m).inf;
            sup=((Sort)m).sup;
            if( sup-inf+1<=50 ){ //
                selection sort of subvector v[inf..sup]
                if( parent!=null ){ send( new Done( parent ) ); }
                else{ System.out.println( java.util.Arrays.toString(v) ); }
                doDelete(); //terminate agent
            }
            else{ //divide-et-impera on subvector v[inf..sup]
                med=(inf+sup)/2;
                Object[] args={control}; //create left child agent
                newActor("s1_"+inf+"-"+med,"actor.mergesort.Sorter",args);
                AID s1aid=
                new AID( "s1_"+inf+"-"+med, AID.ISLOCALNAME );
                Sorter.Sort s1=new Sorter.Sort( s1aid,getAID(),v,inf,med );
                send( s1 );
                create right child agent and initialize it with a Sort msg
                become( MERGING );
            }
        } break;
    case MERGING:
        if( m instanceof Done ){ doneCount ++;
            if(doneCount==2 ){
                merge v[inf..med] with v[med+1..sup] in a sorted sequence
                if( parent!=null ){ send( new Done( parent ) ); }
                else{ System.out.println( java.util.Arrays.toString(v) ); }
                doDelete();
            }
        }
    } //switch
} //handler

public void setup(){
    super.setup();
    Object[] args=getArguments();
    if( args==null || args.length==0 || args.length>1 ){
        throw new RuntimeException( "Sorter wrong arguments." );
    }
    control=(String)args[0]; bind( control ); become( SORTING );
} //setup
} //Sorter
```

Figure 2 – The Sorter actor

As a JADE agent, the Sorter actor class overrides the `setup()` method where, besides invoking `setup()` on the super class, the control name is received as an argument and it is used to bind this actor to the control form. The `setup()` method is also in charge of setting the initial actor state.

The actor-based mergesort application was tested under Concurrent and then executed under Parallel control so as to take advantage of a multi-core architecture. As a lesson learned, due to the synchronization burden in the Actor `newActor()` method, the Parallel execution slightly outperforms the corresponding Concurrent version.

A MODELLING AND SIMULATION EXAMPLE

The following reports an agent-based modelling and simulation experience conducted in JADE about a queueing network termed CSM –Central Station Model– (see Fig. 3). It served to test specifically the usage of the Simulation control structure. The chosen model is representative of a class of realistic models. An adaptation of CSM has been used, for example, for solving by simulation a seaport logistic problem, i.e. optimal assignment of berth slots and cranes to shipping services at a modern marine container terminal (Laganà *et al.*, 2006).

The CSM model is based on K recirculating clients or jobs. Initially the K clients are injected into the reflective station S_0 where they reflect a certain amount of time before re-entering the system. The reflection station makes it possible for all the arrived clients to reflect in parallel.

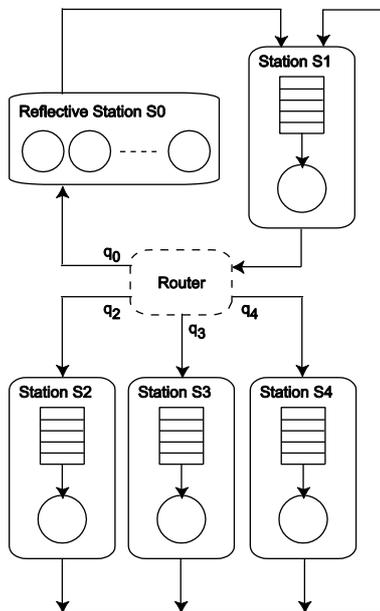


Figure 3 – A CSM model

A client enters the system by arriving to the central station S_1 whose service time is exponentially distributed. After S_1 processing, the client, with certain probabilities (q_0 , q_2 , q_3 , q_4), can go (routing) in input to S_0 , or to one of

the service stations S_2 , S_3 or S_4 . S_2 has an exponentially distributed service time. Station S_3 has a second order hyper-exponential distribution. Station S_4 , finally, uses an Erlang distribution composed of n identically distributed exponentials with the same rate. A client exiting from S_2 , S_3 and S_4 comes back into input to S_1 .

The parameter values of CSM are collected into Table 1. The second order hyper-exp is characterized by the rate of each exponential component (μ_{31} , μ_{32}), and the probability for choosing one distribution or the other (a_{31} , a_{32}). The hyper-exp is configured to reproduce a burst phenomenon, where “silence times” are due to μ_{32} and burst repetitions are due to μ_{31} .

Table 1 – CSM parameter values

Entity	Type	Values
Station S_0	exp	$\mu_0=0.01 \text{ s}^{-1}$
Station S_1	exp	$\mu_1=1 \text{ or } 2.3 \text{ s}^{-1}$
Station S_2	exp	$\mu_2=0.8 \text{ s}^{-1}$
Station S_3	hyper-exp	$\mu_{31}=5 \text{ s}^{-1}$, $\mu_{32}=0.5 \text{ s}^{-1}$, $a_{31}=0.95$, $a_{32}=0.05$
Station S_4	erlang	$n_4=16$, $\mu_4=0.6 \text{ s}^{-1}$
Router	-	$q_0=0.2$, $q_2=0.3$, $q_3=0.3$, $q_4=0.2$
#Circulating clients	-	$K=2, 5, 10, 20, \dots, 100$

Fig. 4 shows an UML class diagram of the CSM model implemented in JADE using actors. AbstractStation defines a generic station and introduces the basic Arrival and Departure messages.

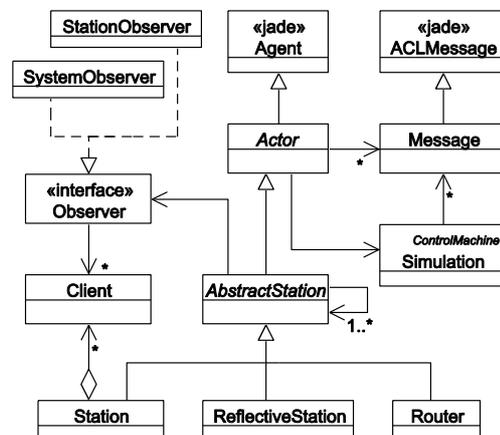


Figure 4 – A CSM model class diagram

Station, ReflectiveStation and Router are concrete heirs of AbstractStation. Station introduces the buffer for waiting of clients which find the server busy. Router and ReflectiveStation have no internal buffering. Stations have an `nextStation` attribute which specifies where a processed client should be directed. Router has an array of next stations each one paired with a probability value. Station initialization, e.g. transmitting the control form, the distribution to be used, the simulation time limit, the identity of the next topological station or array of next stations, rests on the `newActor()/setup()` methods.

Internally, a station uses an Observer for monitoring the occurrence of arrival/departure events. Two particular observers are associated with monitoring a normal station (like S1, S2, S3 or S4) or the entire system through the reflective station (S0). It is worth noting that a client that arrives to S0, actually exits the system. A client which departs from S0, really enters the system. A client object has an attribute for storing the time when it enters the system. When the client exits the system, the current time and the entering time of the client allow to estimate its dwell time in the system, which contributes to defining the response time of the system.

Simulation Experiments

Some experiments were used to estimate by simulation quantitative properties such as the response time (waiting time plus service time spent by a client in a station), throughput (number of clients processed per time unit), etc. of the whole CSM system (emerging properties) and of each single component station S1, S2, S3 and S4, in the two scenarios where the rate μ_1 is 1 or 2.3, and by varying the number of re-circulating clients. Each simulation experiment was executed with a time limit of 3×10^7 time units which guarantees (as it was checked experimentally) the average service time of each station is eventually met. Experiments were carried out on a Win 8, 12GB, Intel Core i7-3770K, 3.50GHz.

System level behavior

Emergent behavior of the system is the result of the interactions and behavior of individual component stations. Bottleneck in some component can affect the whole CSM system.

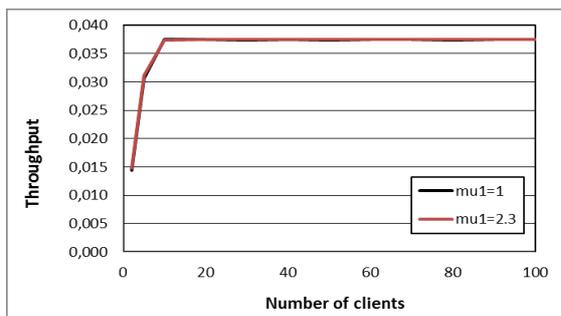


Figure 5 – System throughput vs. the number of clients

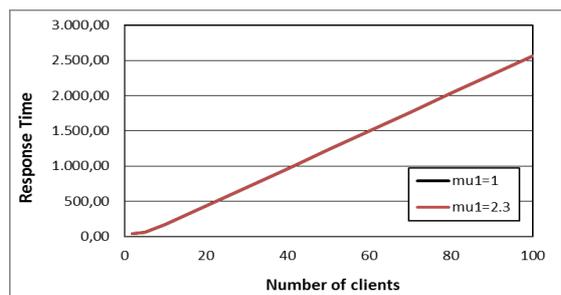


Figure 6 – System response time vs. the number of clients

Figures 5 and 6 show respectively the measured throughput and the response time of the system. As one can see, on the basis of the parameter values in Table 1, no real difference emerges in the two scenarios of $\mu_1=1$ and $\mu_1=2.3$. In addition, from Fig. 5, a saturation occurs as the number of clients becomes greater than or equal to 10. All of this mirrors the fact that, although an increasing number of clients is considered, the system is unable to improve its productivity. Saturation of Fig. 5 is coherent with a sharp augment of the average response time of the system as the number of clients increases. Fig. 6 confirms that clients tend definitely to remain within the system, thus wasting the overall response time. The property is clearly the consequence of some bad-behavior (bottleneck) existing within the “black box” of the system.

Station level behavior

Being the access point to the system, the central station S1 could be naturally a bottleneck for the system. However, as depicted in Fig. 7 its behavior was found to be a “not offending” one for the system.

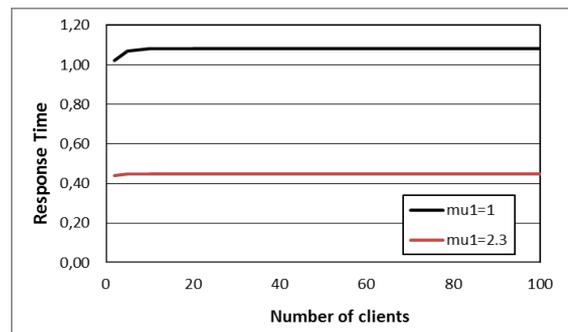


Figure 7 – Central station S1 response time vs. the number of clients

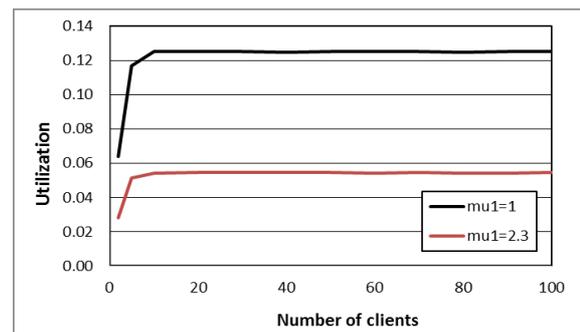


Figure 8 – S1 Utilization vs. the number of clients

For brevity, the throughput and response time of station S1 are not reported but they confirm the achievement of a saturation condition as soon as the number of clients reaches and goes over 10. The utilization of S1 (see Fig. 8) is definitely low and about 0.12 when $\mu_1=1$ and 0.057 when $\mu_1=2.3$. Similarly, the response time of S1 is definitely about 1.1 for $\mu_1=1$ and 0.45 for $\mu_1=2.3$. Bad performance (Fig. 8) when $\mu_1=2.3$ indicates that although the number of clients increases, they are “glued” in some other queue in the system.

A behavior similar to that of S1 was found also for stations S2 and S3, with an eventual value of the utilization being about 0.05 for S2 and 0.007 for S3 (hyper-exponential). Actually, the bad performing station was found to be the S4 (that having an Erlang distribution), as witnessed by Figures from 9 to 11.

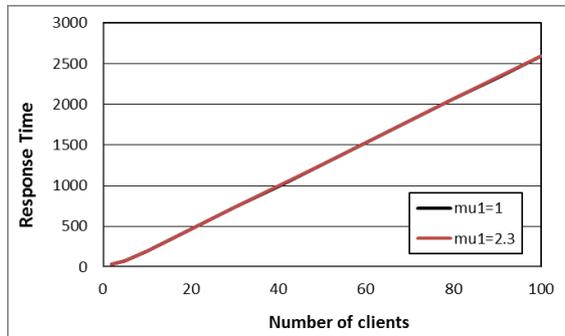


Figure 9 – Response time of S4 vs. the number of clients

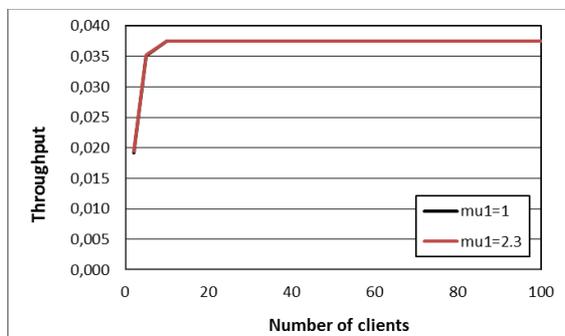


Figure 10 – Throughput of S4 vs. the number of clients

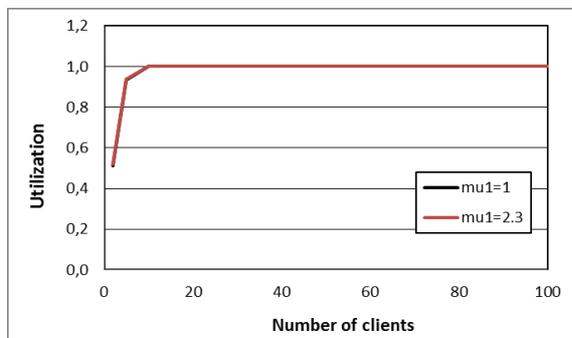


Figure 11 – Utilization of S4 vs. the number of clients

The more clear indicator is the utilization factor of S4 documented in Fig. 11 which at saturation becomes 1, independently from the value of μ_1 , also considering an almost equally distributed routing probability. In other words, S4 is totally engaged with processing clients and further arriving clients await into its queue. On the other hand, the average service time of S4 is 26.66 which is the maximum among the internal stations of the CSM system. Therefore, an S4 service lasts too long thus forcing S1, S2 and S3 to be idle most of the time.

CONCLUSIONS

The JADE based control framework proposed in this paper is open, flexible and useful in the practical case. It

is in current use in an undergraduate course of software engineering for real-time and agent-based systems. Students contributed to the definition of other control forms, e.g. a variant of Realtime where by knowing the duration (worst case) of a message processing, the delivery of a not time constrained message can be delayed in order to better serve time constrained messages.

Prosecution of the research work is geared at

- optimizing the framework implementation
- extending the catalog of reusable control forms
- completing the implementation of distributed control structures, notably by porting to JADE the conservative time synchronization strategies successfully experimented in (Cicirelli *et al.*, 2009, 2011, 2013b, 2014), also in the presence of models with agent mobility.

REFERENCES

- Agha G. 1986. *Actors: A model for concurrent computation in distributed systems*. The MIT Press.
- Bellifemine, F., G. Caire, D. Greenwood (2007). *Developing multi-agent systems with JADE*. John Wiley & Sons.
- Cicirelli, F., A. Furfaro, L. Nigro (2009). An Agent Infrastructure over HLA for distributed simulation of reconfigurable systems and its application to UAV coordination. *Simulation Trans. of the Society for Modelling and Simulation International*, **85**(1):17-32.
- Cicirelli, F., A. Furfaro, A. Giordano, L. Nigro (2011). HLA_ACTOR_REPAST: An approach to distributing Repast models for high-performance simulations. *Simulation Modelling Practice and Theory*, **19**(1):283-300.
- Cicirelli, F., A. Furfaro, L. Nigro, F. Pupo (2013a). Agent methodological layers in Repast Symphony. In *Proc. of ECMS 2013*, pp. 68-74.
- Cicirelli, F. & L. Nigro (2013b). An Agent framework for high performance simulations over multi-core clusters. In *Proc. of 13th AsiaSim 2013*, Springer, Communications in Computer and Information Science (CCIS) series, pp. 49-60.
- Cicirelli, F., A. Giordano, L. Nigro (2014). Efficient environment management for distributed simulation of large-scale situated multi-agent systems. *Concurrency and Computation: Practice and Experience*, to appear.
- Derksen, C., C. Branki, R. Unland (2011). Agent.GUI: A multi-agent based simulation framework. In *Proc. of FedCSIS'11*, pp. 623-630.
- FIPA, Foundation for Intelligent Physical Agents, on-line, <http://www.fipa.org>
- JADE, on-line, <http://jade.tilab.com>
- Gianni, D., G. Loukas, E. Gelenbe (2008). A simulation framework for the investigation of adaptive behaviours in largely populated building evacuation scenarios. *OOAMAS Workshop, AAMAS Conference, Presentation tool*.
- Laganà, D., P. Legato, O. Pisacane, F. Vocaturo (2006). Solving simulation optimization problems on grid computing systems. *Parallel Computing*, **32**(9):688-700.
- Tanenbaum, A.S. & M.V. Steen (2007). *Distributed systems – Principles and paradigms*. 2nd Edition, Pearson Education.
- Wooldridge, M. (2009). *An introduction to multi-agent systems*, 2nd Edition, John Wiley & Sons.

SIMULATING SOCIAL NETWORKS IN SOCIAL MARKETING

Roderick Duncan, Luisa Perez-Mujica, Terry Bossomaier
Charles Sturt University
Bathurst, NSW 2795, Australia
Email: rduncan@csu.edu.au

KEYWORDS

Agent-Based Modelling; Social Networks; Community-Based Social Marketing

ABSTRACT

Community-based social marketing is a relatively new approach in marketing which makes use of social networks within communities to disseminate marketing messages in a information campaign. Agent-based models can be used as a tool for exploring the sensitivity of such campaigns to the structure of social networks within target communities. We develop an agent-based model for a social marketing campaign within surrounding communities for ecotourism services in a wetland. As the revenues from ecotourism are used to fund the rehabilitation efforts in the wetland, the long-term sustainability of the wetland is dependent on the success of the marketing of the ecotourism services. We find that for a small world social structure the success of the social marketing campaign is highly dependent on the number of links between communities. The results suggest that the design of social marketing campaigns may need to take account of social networks between and within communities.

INTRODUCTION

Ecotourism has been a rapidly growing area over the last couple of decades. Thus it generates a stimulus to recover derelict areas which have the potential to create interesting ecological environments. The Winton Wetlands, the subject of this paper, is one such area.

Ecotourist areas need tourists to generate revenue to support their upkeep and further development or rehabilitation. However the tourists' consumption of ecotourism services often causes damage to the same environmental services the tourists are there to enjoy, through water or noise pollution, wastes and simple disruption of the environment. Ecotourists also demand non-ecological facilities, from the most basic needs through to coffee bars, restaurants, coffee shops, information kiosks and helpful rangers to provide on-the-spot information within areas of high ecological and biodiversity value.

Getting the right balance is difficult between the desire for ecotourists to have access to areas of high ecological value, the financial need in terms of long-term sustainability for the revenues that come with those ecotourists and the protection of those ecological val-

ues in the areas which make them so attractive to ecotourists. This paper argues that simulation, specifically Agent Based Modelling, is a good way to explore the strategic options available to natural resource managers. In fact the simulation suggests a non-intuitive finding as to the requirements of the most effective marketing.

This balance has been explored using a system of differential equations elsewhere [LPCZ07]. But such a systems-dynamics like approach is unable to capture the diversity of behaviours in a complex system, which in this case is a combination of an ecological system of a wetland and a human system of a marketing effort in an attempt to attract ecotourists. Any attempt to model such as system must include details of both the human social networks and the community-based marketing effort and the details of the wetland itself.

Social Networks

One exciting trend of the last two decades has been the understanding of networks and network structure, summarised in Paul Ormerod's recent book [Orm13]. Two formative developments were the introduction of *small-world* networks by Watts and Strogatz [Wat99] and *scale-free* networks by Barabási [Bar02]. Network studies have spread to many areas, from the study of pandemics [GKLM06] to the appearance of small-world structures in language [FiCS01]. Numerous methods now exist for eliciting social network structure [KW06].

Small world and scale-free networks have a property in-common, their low distances between nodes, where distance is defined as the number of hops between connected nodes to get from one node to another. In small-world networks these short paths arise from short cuts, a bit like the snakes and ladders of the popular children's board game.

There are many ways of defining a small-world network structure. The one we use here is to start with a regular network with each node connected to a small number of neighbours. Then links are broken at random, forming a new link to some other random location. The number of such cuts, N_c determines how short the average path length becomes. Quite small values of N_c lead to low path lengths, as encapsulated by Milgram's famous *six degrees of separation* experiment [Mil67].

Scale-free networks achieve this in a different way, through having highly connected *hubs*. The quintessen-

tial example would be the airport network. To fly from some small airport in Russia to another in the USA, would usually involve flying in and out of hubs, say Moscow and JFK, both of which connect to many airports.

Community-based social marketing

Social marketing is a subdiscipline of marketing which seeks to influence human behaviour to achieve some social good or to ameliorate some social problem using marketing tools [KZ71], [KL11]. Social marketing has been proven of use in many public campaigns to address problems such as drink driving rates, teen smoking rates or low use of mosquito nets to combat malaria [KL11].

In contrast to more traditional public information campaigns which typically make use of mass media, community-based social marketing (CBSM) uses existing social networks between tourists to disseminate marketing messages [MM00], [MMS11]. Such campaigns may be less expensive to run without the high cost of purchasing space within the mass media but are dependent on social networks for their dissemination.

In a typical CBSM campaign, the first consumers contacted by the marketers are encouraged not only to engage themselves to the project, but also to engage with their friends and family to encourage their social network to also join the campaign. In an environment with a small world social structure, the problem may arise that, since consumers tend to be linked to people like themselves, the members of those social networks will already have been exposed to the marketing message. The difficulty in the social marketing campaign might be engaging with those people who are socially very distant.

The logic of CBSM is similar to that of a contagion. Those consumers infected with the marketing message are encouraged to infect others, but in this case the infection might be a desirable outcome. As with a contagion model, the key vectors might be those customers who cross large social divides to infect entire new communities and where a first response to an outbreak is to limit travel between populations [GKLM06].

Stylized facts about the Winton Wetland

The Winton Wetlands is a restored wetland site, near the regional township of Benalla, in Victoria, Australia. The site has been subjected to a range of land-use and hydrological regimes, since the establishment of agricultural activity in the area. The land in the surrounding catchment was initially developed for agricultural use, and the wetlands were dammed, to form an irrigation reservoir known as Lake Mokoan. After more than three decades as a water-supply reservoir, the dam was decommissioned and the wetlands were drained to restore the site to as close to the original state as possible. The Winton Wetlands are located inside what is considered Victorias High Country Tourism Strategy. A review of the draft of the tourism strategy indicates that the Winton Wetlands would be a strong fit with three of

the regions five strategy pillars: cycling, nature-based experiences and cultural heritage [LeaS⁺99].

The area of the wetlands is approximately 9,000 hectares, of which half consists of the wetland itself, and the remainder consists of a mix between native woodland and altered grassland, which are also home to numerous native fauna and migratory bird populations. The Winton Wetland rehabilitation is the largest wetland restoration project in the southern hemisphere, with the aim of restoring biodiversity values and reintroducing native species within the wetland [Aut12]. There are three communities within a 50 km radius to the wetland; the rural cities of Glenrowan, Benalla and Wangaratta with a population of 963, 13,643 and 26,816 people respectively [oS11].

Under the current plan for the wetlands [LeaS⁺99], the wetland managers will invest A\$7 million initially and then attract a mix of A\$25 million in government funding and a further A\$25 million in private capital to develop the ecotourism infrastructure in the wetlands. This infrastructure will include a range of accommodation and cafes, as well as an education centre, tree top walk, boardwalks, boat ramps, bike and walking trails and staging for events. The Victorian state government has pledged A\$20 million for the wetlands project. The initial local market within 50 kilometers of the wetlands has a population of 67,000, including the three communities mentioned above, and is estimated to provide 140,000 repeat tourist visits per year out of the estimated 340,000 total visits to the wetland per year. Engagement with the local community around the wetlands is thus a crucial component of the marketing strategy envisaged by the wetlands management. It is this local marketing effort which we will simulate in this paper.

AGENT-BASED MODEL OF A WETLAND ECOTOURISM MARKETING CAMPAIGN

The simulation model examines the effectiveness of small-world networks in coupling across communities. The framework of the model has been described elsewhere [PMBD⁺13], [PMDB14]. However in that paper tourists shared information about the wetlands with other tourists in a random fashion, whereas herein tourists are now assumed to share information about the wetlands through their fixed social links, which are determined at the start of each simulation.

Social network analyses have shown that typical social networks tend to have a particular structure. Our previous paper had a completely random network structure, but human societies generally to have small world structure where people tend to be linked to other people with whom they share links. The typical small world network structure tends to have lots of short links, where people are tightly connected to those close to them (in terms of geography, family structure, socio-demographics, etc) and only a few long links where people are connected across large social gaps. The social structure of the community is assumed to be a small world. Within each village, the villagers share many

social links however across villages the links are rare or non-existent. The next section provides a general description with technical details in the following section.

Overview of the simulation model

There are two main types of agents in the model: tourists and rangers. The tourists - or potential tourists - are located in the surrounding three villages and may choose to visit the wetland as a recreational experience. The rangers interact with the tourists visiting the wetlands, gather resources from the tourists and expend those resources rebuilding and rehabilitating the wetlands.

For simplicity the wetland has only two broad features of interest to tourists: ecology and infrastructure. During a tourist’s recreational visit to the wetland area, the tourists interact with the wetland and update their experience based on the levels of ecology and infrastructure which they encounter - modified by the tourist’s individual preferences. As tourists interact with the wetland, they consume the ecology and infrastructure of the wetland, which the rangers then repair using resources generated from interactions with the tourists. The alternative tourist activity is labelled the “beach”, which represents all other forms of recreation the tourists may engage in. Tourists choose to visit the wetland or the beach based on their past experience and on information about the wetland which they receive through their social network.

The model in this paper features tourists from three townships, with progressively lower levels of participation and stakeholder involvement in the Winton Wetlands. These are the towns of Benalla, Wangaratta and Yarrowonga. Residents in these communities are strongly connected to the people within their own village and sparsely connected to the other villages. Each village is a random small-world network in its own right, but then we break intra-village lengths, and add inter-village links. This reflects network structures within and among communities as detected, for example, by Twitter [BFJ13]. This method is a shortcut to generating networks with community structure directly, as for example, in Salaberry [SZM13]. These sparse, long links make use of homophily between tourist [Jac08], [Bur27], [MSLC01], [LM54] where tourists are more likely to be linked to tourists in other communities who have similar preferences to their own.

We simulate the transfer of information about the wetland – word of mouth [GLM01] – through the social links which exist within the community around the wetland. The local community of three villages differs according to how much information they initially have about the wetland. The success of the wetland depends on the positive experiences tourists have at the wetland and how effectively positive information about the wetland is communicated through the community. There is a small chance each time step that a tourist will exchange wetland experience information with other tourists through the social network for that

tourist. The new value for experience is assumed to be a simple average of the tourist’s own experience and the experiences of the tourists to whom the agent has a social link.

The long-term survival of the wetland depends on the revenues raised from ecotourism activities within the wetland. These ecotourism revenues pay for the ecological restoration activities and the infrastructure maintenance conducted by the wetland rangers. The tourism infrastructure within the wetlands is assumed to depreciate over time, so the maintenance of the wetland depends on a constant stream of tourists. These same tourists however also damage the ecology of the wetland and the balance between ecological harm caused by the tourists and revenue raised by the tourists is an important feature of this sort of model.

Technical Details

The model is implemented in Netlogo [Wil99]. Figure 1 shows a screen shot of the environment. The basic model has been described in [PMD14]. It comprises an environment, a set of agents and a connectivity graph. Values for the parameters are given in Table .

The Environment

The environment consists of the wetlands, the beach and the three communities. Each is represented as a type of patch in Netlogo. No specific activities are modelled in the beach, however, so it is essentially a boundary condition.

The communities are homogeneous, each consisting of Z_v patches, while the beach has Z_b patches. The wetland has Z_w patches, which are initially randomly seeded with a level of infrastructure Z_{wi} and a level of ecology Z_{we} . These values are updated continuously throughout the simulation as the agents interact with the patch on which they are located. The wetland environment is the heart of the model, and more structure is being continually added, as in the companion paper in this conference [PMBD14].

The Agents

There are two broad types of agents in the model used for this paper: N_r rangers and N_t tourists. Rangers are only located in the wetlands, and tourists are initially only located in the communities. The rangers are further divided into two classes: builders who repair the infrastructure and ecologists who the repair ecology for the patch on which they are located. Each ranger is initially allocated a level of resources, which the ranger will use to repair the infrastructure or ecology - depending on the type of ranger - of the patch on which the ranger is located at the cost of the level of resources which the ranger possesses. At each time step, the ranger roams to an adjacent patch of the wetland and increments the infrastructure or ecology value of the patch at the cost of 1 unit of resources the ranger possesses.

The ranger agents will continue incrementing the infrastructure or ecology of the wetland patches until the patch reaches its maximum level for that quality, or the ranger runs out of resources. The only means for rangers to increase resources is to receive them from a tourist in the wetlands. At any time step when a tourist and a ranger are co-located on a patch of the wetlands, the tourist will make a payment of p resources to the ranger increasing the ranger’s resource level. This payment mechanism is intended to mimic the choice of the wetlands managers under the management plan [LeaS⁺99] that entry fees were not the appropriate mechanism of revenue raising for the wetland but rather fees for particular services would be preferable.

The tourists are initially located in three communities with social networks as described in the next subsection. Each tourist is randomly seeded with a wetland experience which is then updated through the simulation. The initial values for the wetland experience are set at different average levels for the three communities to represent the degree to which the communities are currently integrated with the wetland. The Benalla community is closely tied in with the Winton Wetlands, however the Wangaratta community already has existing water recreation facilities along the Ovens River.

At each time step tourists who are currently not in the wetlands have a 0.6% chance to consider a visit to the wetlands. The probability was chosen to mimic the 140,000 repeat visits for the 67,000 residents in the local community of the wetlands under the management plan. However the tourist will only choose a wetland visit if the tourist’s experience of the wetland, based on past visits and word-of-mouth, exceeds the expected experience of a visit to all alternative choices of recreation. If the tourist does choose to visit the wetland, the tourist will roam the wetland patches at each time step until the tourist hits the boundary of the wetland and is returned to its community. Thus only a small fraction of the tourist agents are interacting with the wetland at any point in time.

At each time step in the wetland, the tourist updates the wetland experience value. The updated value is a weighted average of the tourist’s existing wetland experience, δ_w , and the wetland experience for the current time step, $1 - \delta_w$. Calculating the current time step’s experience for the tourist depends on the tourist’s preferences for infrastructure and ecology and the levels of infrastructure and ecology for the wetlands patch on which the tourist is located. The preference vector for each tourist is \bar{e} with elements, e_i in the range [0..1] of features in the wetland which they value. The preference vector for each tourist is determined randomly and denotes the relative importance of each feature of the wetland - infrastructure and ecology - in the calculation of the tourist’s wetland experience.

The Tourist Social Network

The network of social links between people in each community is dense. But a small number, η_x , of links

Parameter	Symbol	Value
Wetland patches	Z_w	624
Wetland ecology level	Z_{we}	[0..10]
Wetland infrastructure level	Z_{wi}	[0..10]
Village patches	Z_v	1872
Beach patches	Z_b	52
Tourists per village	N_t	50
Wetland experience decay	δ_w	0.9
Tourist preferences	\bar{e}	[0..1]
Number of small world links	η_x	[0..10]
Prob of intra-community links	η_l	0.7
Network influence	α	0.5
Tourist payment	p	[30..50]

TABLE I: Model parameters and their values

are broken and connected to an agent in another community. This connection is biased in favour of people, j, k with similar tastes in other communities (homophily) – defined as the Euclidean distance, h_{jk} , between their preference vectors.

$$h_{jk} = \sqrt{\sum_i (e_i^{(k)} - e_i^{(j)})^2} \quad (1)$$

The simulation has three villages of tourists, and each village has N_t tourists. (The representation in Netlogo appears in Figure 1). Each tourist has η_l percent probability of being linked to another tourist in the same village. A small number of social links, η_x within villages were changed to long social cross links across villages.

To form the cross-links, a tourist in one village is selected. The homophily, defined by Eqn 1 is then calculated for each tourist outside the selected tourist’s community is computed. The link is then made to the tourist outside the selected tourists’s with the greatest homophily.

Simulation dynamics

Each simulation begins with a random assignment of tourists to patches in the villages. Each time step in the simulation is intended to represent a tourist day in the wetlands. At each time step each tourist and each ranger within the wetlands moves to an adjacent patch in the wetlands. When outside the wetland, in the village, each tourist, j , updates his/her estimate of the wetland quality, $q_j(t - 1)$ at time $t - 1$ according to any changes of opinion by their links in their social network, S according to eqn 2

$$q_j(t) = (1 - \alpha)q_j(t - 1) + \alpha \sum_{i \in S} q_i(t - 1) \quad (2)$$

The impact of cross-community links was investigated by running 100 simulations from random starting conditions for different values of the payment from tourist to ranger.

RESULTS FROM THE SIMULATIONS

Figure 1 shows a standard run of the simulation. The green area is the wetland, the brown the three villages, and yellow the beach. The agents represented in the simulation are the rangers located in the wetland and the potential located in the villages with the black lines breaking up the tourists into the three villages. The grey lines between tourists represents the social links between them, which also are the mechanism of diffusion of information within and between the villages.

Figure 1 shows the Netlogo map.

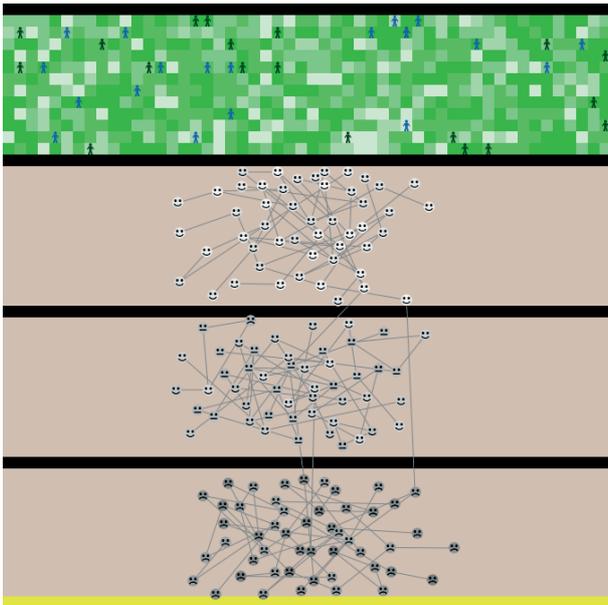


Fig. 1. The World Map in Netlogo

In the simulations we modeled the effect of adding a limited number of long social links - between the three villages. The model is run 100 times for different numbers of long links between the villages, while the values of tourist payments η were allowed to vary between 30 and 50. The success of the wetland in each run is measured by the final level of infrastructure for the simulation, which is a measure of the long-term viability of the wetland.

Figure 2 shows how the park benefits from increasing number of links between communities, where each line represents the average final value of infrastructure varying the number of long links but holding the payment between rangers and tourists constant. The model displays a sharp change of behavior around the value of 40 for the payment.

We found that adding more long links greatly increased the probability of the wetland having a successful CBSM campaign as long as the payment between the rangers and the tourists is at the right level. In other words for CBSM it matters how many links tourists share, but also what types of links.

The model is particularly sensitive to the level of payment made by the tourists as shown by Figure 1. For a payment level of 35 and below, even a high level of social links between communities could not save the

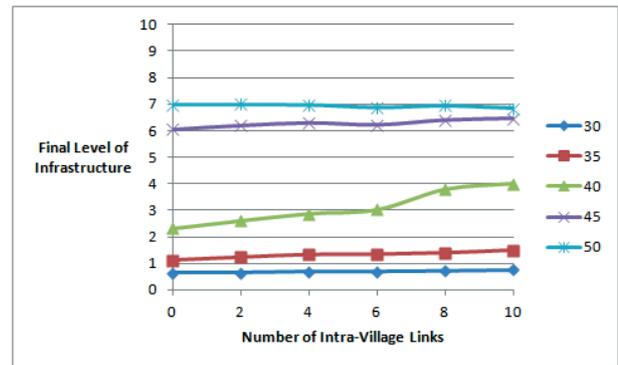


Fig. 2. Plot of Infrastructure Level as a Function of Different Numbers of Inter-Community Links

wetlands, as the rangers did not receive enough revenues to maintain the attractiveness of the wetlands for ecotourists. For a payment level of 45 and above, the revenues were sufficient to retain tourists without any use of social marketing.

DISCUSSION AND CONCLUSION

Knowing and correctly using the social structure within the community might be an important feature of a social marketing campaign. A CBSM campaign may not be successful under a small world social structure if the tourists reached by the campaign are highly linked with tourists who have already been communicated with, as all tourists are tightly linked locally. In a small world social structure this feature is likely to be true.

This result suggests that identifying tourists who are likely to have long social links may be a successful strategy for a CBSM campaign, as might identifying tourists who are not like the usual tourist in some feature, as these tourists may allow the marketing campaign to reach out to communities who would not have heard much about the wetland before the campaign.

The sharp difference in behaviour in the simulations around the critical payment parameter value provide a warning about the types of marketing campaigns which might be used by the wetland managers. One possible marketing response to low ecotourist visits would be to cut the price of visits, so raising the volume of ecotourism and perhaps the revenues raised by the wetland. However in the case of ecotourism, an increase in volume of ecotourism is also an increase in damage to the wetland which will have to be repaired. A cut in price may make the financial situation of the wetland far worse with high levels of ecological restoration needed for the wetland with little new revenue. Wetland managers may be far better off using other sorts of marketing campaigns.

AUTHOR BIOGRAPHIES

RODERICK DUNCAN is a senior lecturer in economics and finance at Charles Sturt University. His research interests include international trade, development economics, resource and environmental economics and applied microeconomic analysis.

TERRY BOSSOMAIER is professor of complex systems at Charles Sturt University, specialising in complex systems and their simulation. His work ranges from Agent Based Models of social systems to the study of critical phenomena. He is interested in the behaviour of information theoretic quantities around phase transitions and the prospects of being able to understand herding behaviour or bubbles and crashes. He is the author/editor of five books.

LUISA PEREZ-MUJICA is a doctoral student at Charles Sturt University, but is originally from Mexico. Her project centres around the study of environmental restoration of a wetlands area, combining stakeholder interviews and other inputs with agent based simulation of wetland tourist development.

REFERENCES

- [Aut12] Authority, G. B. C. M. Lake mokoan decommissioning plan. <http://www.gbcma.vic.gov.au>, 2012.
- [Bar02] Barabási, A.-L. *Linked*. Perseus, Massachusetts, 2002.
- [BFJ13] Bryden, J., Funk, S., and Jansen, V. Word usage mirrors community structure in the online social network twitter. *EPJ Data Science*, 2(1):3, 2013.
- [Bur27] Burton, R. *The Anatomy of Melancholy*. Farrar and Rinehard, New York, 1927.
- [FiCS01] Cancho, R.Ferrer i and Solé, R. V. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268:2261–2265, 2001.
- [GKLM06] Germann, T., Kadau, K., Longini, I., and Macken, C. Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006.
- [GLM01] Goldernberg, J., Libal, B., and Muller, E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.
- [Jac08] Jackson, M. *Social and Economic Networks*. Princeton University Press, 2008.
- [KL11] Kotler, P. and Lee, N. *Social Marketing: Influencing Behaviours for Good*. SAGE 4th Ed., 2011.
- [KW06] Kossinets, G. and Watts, D. J. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
- [KZ71] Kotler, P. and Zaltman, G. Social marketing: An approach to planned social change. *Journ. Marketing*, 35:3–12, 1971.
- [LeaS⁺99] Lethlean, T. C., al, M.et , Sanmor, , DesignFlow, , QS, P., and Arup, . The win-ton wetlands at benalla master plan, 1999.
- [LM54] Lazarsfeld, P. and Merton, R. Friendship as a social process. In Berger, M., editor, *Freedom and Control in Modern Society*. Van Nostrand, New York, 1954.
- [LPCZ07] Lacitignola, D., Petrosillo, I., Cataldi, M., and Zurlini, G. Modeling socio-ecological tourism-based systems for sustainability. *Ecol. Modelling*, 206:191–204, 2007.
- [Mil67] Milgram, S. The small world problem. *Psychol. Today*, 2:60–67, 1967.
- [MM00] McKenzie-Mohr, D. Promoting sustainable behavior: An introduction to community-based social marketing. *J. Social Issues*, 56(6):543–554, 2000.
- [MMS11] McKenzie-Mohr, D. and Smith, W. *Fostering Sustainable Behavior: An Introduction to Community-Based Social Marketing*. New Society Publishers, Gabriola Island, 3rd. Ed., 2011.
- [MSLC01] McPherson, M., Smith-Lowin, L., and Cook, J. Birds of a feather: Homophily in social networks. *Ann. Rev. Sociology*, 27:415–444, 2001.
- [Orm13] Ormerod, P. *Positive Linking*. Faber and Faber, 2013.
- [oS11] Statistics, A. B.of . *2011 Census Community Profiles*. Canberra, 2011.
- [PMBD⁺13] Perez-Mujica, L., Bossomaier, T., Duncan, R., Rawluk, A., Finlayson, C., and Howard, J. Developing a sustainability assessment tool for socio-environmental systems: a case study of systems simulation and participatory modeling. In *Proc. Modeling and Applied Simulation, Athens, Greece*, 2013.
- [PMBD14] Perez-Mujica, L., Bossomaier, T., and Duncan, R. Developing a sustainability assessment tool for socio-environmental systems: a case study of systems simulation and participatory modelling. In *Proc. European Conference on Modeling and Simulationd, Brescia, Italy*, 2014.
- [PMD14] Perez-Mujica, L., Duncan, R., and Bossomaier, T. Using agent-based models to design social marketing campaigns. *Australasian Marketing Journal*, 2014.
- [SZM13] Sallaberry, A., Zaidi, F., and Melanon, G. Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, 3(3):597–609, 2013.
- [Wat99] Watts, D. J. *Small Worlds*. Princeton University Press, 1999.
- [Wil99] Wilenski, U. Netlogo, 1999.

Simulation of Complex Systems and Methodologies

Simulation of a muon based monitoring system

G. Bonomi, A. Donzella, M. Subieta, A. Zenoni
Department of Mechanical and Industrial Engineering
University of Brescia
I-25123, Brescia, Italy
Email: germano.bonomi@unibs.it

KEYWORDS

Positioning and alignment; Computer modeling and simulation; Cosmic rays

ABSTRACT

In recent years cosmic rays have been suggested for civil applications. In this work a specific simulation to assess the feasibility of a muon based monitoring system for historical buildings static stability is presented. In particular the monitoring of the wooden vaulted roof of the “Palazzo della Loggia” of the City of Brescia will be described and the corresponding results will be presented.

INTRODUCTION

Cosmic rays are particles that hit the Earth surface with a rate of about 10000 per minute and per squared meter. Most of them are “muons” and are generated in reactions between the primary rays (coming from the sun and from other extrasolar sources) and the Earth atmosphere. They have a mean energy of about 3-4 GeV and they are able to penetrate even several meters of rock [1]. They have been used to investigate the intimate structure of matter and are often used in particle and nuclear physics to test and calibrate detectors. Recently various research groups have suggested to exploit them for civil and security applications [2]-[11]. Simulation of the interaction of muons with matter are clearly essential to project and design instruments and systems for such civil and security applications. In the present work a stability monitoring system for historical buildings based on the detection of cosmic ray muons will be presented. In particular the results of the simulation of a specific case, namely the study of static anomalies of the wooden vaulted roof of the “Palazzo della Loggia” of the City of Brescia will be described.

INTERACTION OF MUONS WITH MATTER

When a muon travels through a given material it is slowed down and it is deviated from its original trajectory. The mean stopping power for high-energy muons in matter can be described by $\langle -dE/dx \rangle = a(E) + b(E)E$, where $a(E)$ is the electronic stopping power and $b(E)$ is the energy-scaled contribution from radiative processes (bremsstrahlung, pair production, and photonuclear interactions). $a(E)$ and $b(E)$ are both slowly-

varying functions of the muon energy E where radiative effects are important. High energy muons can easily cross tens of cm of iron and tens of meters of rock. More information and tables can be found elsewhere [12]. For what concerns the deviation from the incoming path of the muons when crossing a given material, the underlying physics is the Multiple Coulomb Scattering (MCS) [13] [14]. The deviation angle, projected on a plane, has approximately a Gaussian distribution with zero mean value and root mean square σ that depends on radiation length X_0 and thickness x of the material and on the inverse of the muon momentum p according to the well-known formulae:

$$\sigma = \frac{13.6 \text{ MeV}}{\beta pc} \sqrt{\frac{x}{X_0}} [1 + 0.038 \log(x/X_0)] \approx \frac{13.6 \text{ MeV}/c}{p} \quad (1)$$

$$X_0 = \frac{716.5 \text{ (g/cm}^2\text{)}}{\rho} \frac{A}{Z(Z+1) \log(287/\sqrt{Z})} \quad (2)$$

where ρ , Z and A are the density, atomic number and mass number of the material, respectively. As an example, for muons of 1 GeV/c momentum traversing a 10 cm thickness, σ is 14 mrad for aluminum, 35 mrad for iron, 64 mrad for lead and 86 mrad for uranium.

THE SIMULATION TOOL

All the physics about the interaction of muons with matter is included in a simulation package developed at CERN and called GEANT4 [15]. Indeed GEANT4 is a toolkit for simulating the passage of particles through matter. It includes a complete range of functionality including tracking, geometry, physics models and hits. It has been designed and constructed to expose the physics models utilised, to handle complex geometries, and to enable its easy adaptation for optimal use in different sets of applications. The toolkit is the result of a worldwide collaboration of physicists and software engineers. It has been created exploiting software engineering and object-oriented technology and implemented in the C++ programming language. It is being used in applications in particle physics, nuclear physics, accelerator design, space engineering and medical physics [15]. It is the most complete, reliable and basically the *de facto* statutory software toolkit

for this kind of simulations. Recently a new interface has been implemented that allows users to record the simulation output and to define and handle the geometry via the ROOT package [16], this new system being called GEANT4 VMC [17]. ROOT is a CERN software package specifically developed for particle physics, even though it is now used in many other fields. It has been written in C++ and it now contains all the tools required for data analysis.

THE IDEA

In particle and nuclear physics, muons are often used to “calibrate” the experimental apparatuses, that is to measure the relative position of different detectors with respect one to each other. Recently our research group proposed to use a similar technique for civil applications such as the mechanical monitoring of an industrial press [7]. Since muons are like (almost) straight lines and they can easily cross floors and walls of buildings, a new application for the stability monitoring of historical buildings it is now proposed and studied [18] [19]. The main component of the suggested monitoring system is the “muon telescope”, shown schematically in fig. 1(a). It is composed by a set of three muon detector modules supported by an appropriate mechanical structure and axially aligned at distance of 50 cm one from the other. Each module is composed by two orthogonal layers of 120 scintillating optical fibers with $3\text{ mm} \times 3\text{ mm}$ cross section and 400 mm length, as shown in fig. 1(b).

The two planes of orthogonal scintillating fibers provide the measurement of the crossing position of an incident muon in the x and y coordinates, with a pitch of 3 mm. Considering a flat detection efficiency over the entire surface of the scintillating fiber, the expected spatial resolution on the hit coordinate is about 0.9 mm.

The “muon telescope” is mechanically fixed to a structural element of the building, that constitutes the reference system, with its axis aligned in the direction corresponding to the part of the structure whose displacements should be monitored. A fourth muon detector module, with the same geometry and structure of the previous ones, is positioned as “muon target” on the point to be monitored.

Thanks to their high penetrability, cosmic ray muons are able to cross the system of four detectors as well as the interposed building structures. In this way it is possible to continuously monitor the horizontal displacements of the “muon target” relative to the “muon telescope” fixed on the masonry structure of the building.

Indeed, the trajectory of a cosmic ray muon crossing the system of four detectors can be extrapolated from the “muon telescope” to the plane of the “muon target” detector, in the hypothesis that it is a perfect straight line. The difference between the muon crossing point on the “muon target” and the extrapolated one from the “muon telescope” allows the position of the “muon target” relative to the “muon telescope” to

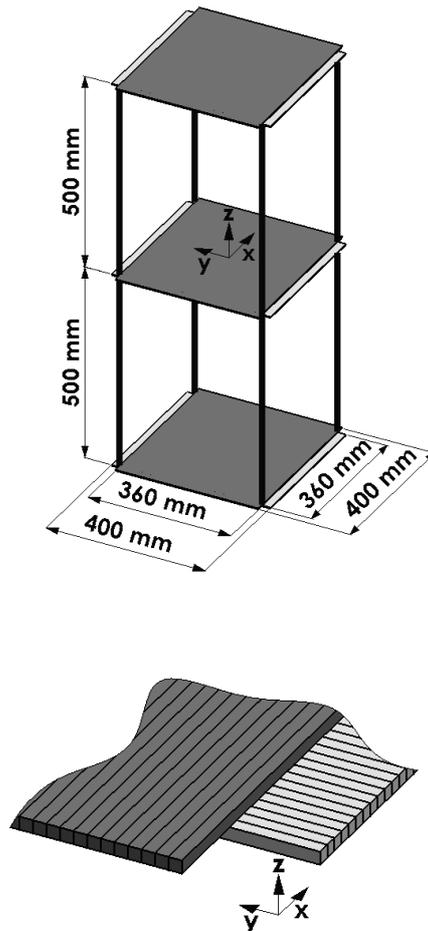


Fig. 1. (a) Structure of the “muon telescope” formed by three muon detector modules axially aligned at a distance of 50 cm each other. (b) Sensitive volume of the muon detector module formed by two orthogonal layers of 120 scintillating fibers $3\text{ mm} \times 3\text{ mm}$ cross section and 400 mm length.

be measured. Possible displacements of the position of the “muon target” relative to a reference position previously determined can be inferred.

In crossing the interposed materials, the trajectories of cosmic ray muons suffer multiple scattering angular deviations. At fixed momentum and at low deviation angles, these deviations follow a Gaussian law with variance depending on the inverse square of the muon momentum [13] [14]. Being these stochastic effects largely dominant over intrinsic detector resolution and geometrical conditions, statistical distributions of the difference between measured crossing coordinates in the “muon target” and the predicted crossing coordinates determined by extrapolation from the “muon telescope” are therefore necessary, in order to reduce the stochastic effects by statistical inference methods. As shown in [7], efficient unbiased estimators of the systematic displacement can be extracted from these distributions. To test the validity of the proposed method a Monte Carlo simulation based on GEANT4 has been



Fig. 2. The “Palazzo della Loggia” of the town of Brescia (1574).

performed for a specific case, that is the static anomalies of the wooden vaulted roof of the “Palazzo della Loggia” of the City of Brescia.

APPLICATION TO A SPECIFIC CASE: “PALAZZO DELLA LOGGIA”

Since its completion in 1574, the “Palazzo della Loggia” (see fig. 2, seat of the municipal hall in the town of Brescia, has cumulated a long sequence of injuries, transformations, repairing interventions, some of which have generated considerable problems of structural stability of the building. The grandiose wooden vaulted roof was completely reconstructed in 1914, with the same architectural shape and construction techniques of the original one, destroyed by a fire one year after the completion of the building. The shape of the dome is like an upside down ship which reaches in elevation a maximum of 16 m, having the planar rectangular sides of about 25 and 50 m respectively. The structural architecture of the vault consists of principal truss wooden arches and simple secondary arches, both connected at the top by a truss made wooden beam.

Immediately after its construction, the present wooden vaulted roof structure exhibited a progressive deformation of the longitudinal top beam and of the key points of the connected arches. The progressive deflection of the top beam was measured to be 190 mm in 1923, 520 mm in 1945, 800 mm in 1980 and it is visible on the top of the roof in fig. 2. Starting from 1990, a systematic campaign of investigation and monitoring of the different stability problems of the Palace (in the following “*the measurement campaign*”) has been committed by the Brescia municipality to the “*Centro di studio e ricerca per la conservazione e il recupero dei beni architettonici e ambientali dell’Università di Brescia*” [20], [21]. In particular, the progressive deformations of the principal arches of the wooden vault have been studied with a specifically designed mechanical measurement system. A progressive collapse of the wooden structure of the arch of about 1 mm per year has been measured.

SIMULATIONS AND RESULTS

The features and expected performances of the proposed measurement system were studied by Monte Carlo simulations using the GEANT4 package [15] above described. A cosmic ray muon generator based on experimental data was implemented in the code in order to simulate as realistically as possible the momentum, the angular distribution and the charge composition of the cosmic ray radiation at the sea level [22].

In order to study the performances of the proposed monitoring system, the structure and composing materials of the “muon telescope” and “muon target” were modeled as well as the relevant structures of the “Palazzo della Loggia” building.

Three configurations were considered: the first with the “muon target” located 0.50 m above the wooden ceiling of the “Salone Vanvitelliano” (this position will be pointed as “P1” in the following), at the first floor of the Palace; the second with the “muon target” located 5.8 m above the wooden ceiling (“P2”); the third with the “muon target” located 10.0 m above the wooden ceiling (“P3”). In the three different conditions the “muon telescope” was located on the vertical of the corresponding “muon target”, 3.0 m below the wooden ceiling. The ceiling of the large “Salone Vanvitelliano” was modeled as a bulky 15.0 cm thick wooden layer.

A. Position measurement uncertainty of the stability monitoring system versus data taking time

Simulation campaigns of populations of cosmic ray muons crossing the measurement system were performed for the three configurations described above. The distributions of the differences Δx and Δy between the crossing point coordinates measured by the “muon target” and the crossing point coordinates extrapolated from the “muon telescope” were calculated.

In figs. 3, sample distributions Δx for the three configurations are shown, for an elapsed data taking time of 15 days, corresponding to about 31.7×10^6 cosmic ray muons crossing the “muon target” surface at the rate of 170 muons/(s m²) [1]. The number of events in each distribution is coherent with the number of cosmic ray muons entering the geometrical acceptance of the measurement system, which depends on the surfaces of the “muon target” and of the lowest “muon telescope” modules and on their distance. The Δy distributions are not shown, since they are statistically identical to the Δx distributions.

As the “muon target” and the “muon telescope” are exactly coaxial in the simulation, the Δx distributions are symmetric and centered at zero. The shape of the distributions exhibits a central narrow peak with very long tails on both sides. This shape is due both to the intrinsic uncertainty of the “muon telescope” in measuring the direction of the cosmic ray muon and to multiple scattering angular deviations of the muon trajectories traversing the interposed materials. The latter effect dominates for large distances of the “muon target” from the “muon telescope”.

The long tails of the distributions are due in part to

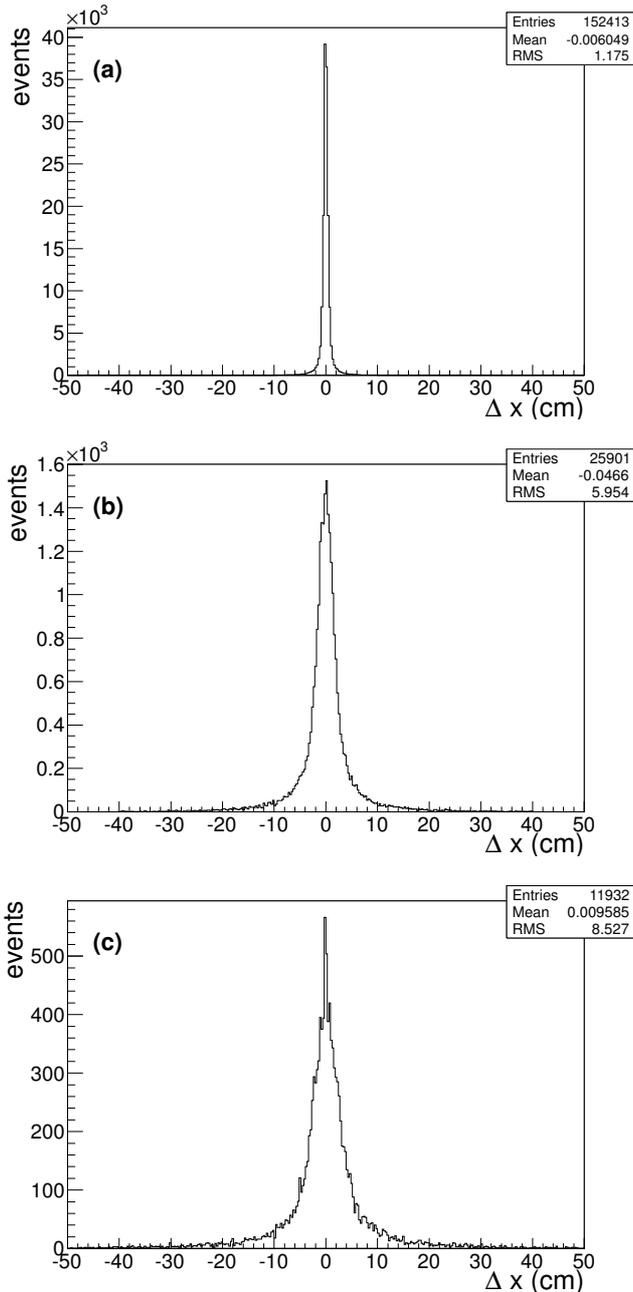


Fig. 3. Distributions of the differences Δx between the crossing point coordinates measured on the “muon target” in position P1 (a), P2 (b) and P3 (c), and the crossing point coordinates of the extrapolated muon trajectory measured by the “muon telescope”, in 15 days data taking time.

low momentum muons, suffering larger deviations, and, in part, to spurious events corresponding to emission of delta rays, most of which can be discarded with a more refined data analysis. At present, the only selection applied to these bad quality events is an arbitrary cut of both tails in the three distributions, discarding about 1 % of the total events.

The mean value of the sample distributions represents an unbiased estimator of the position of the “muon target” relative to the “muon telescope” axis. The root mean square of the sample distribution represents the uncertainty in the measurement of the po-

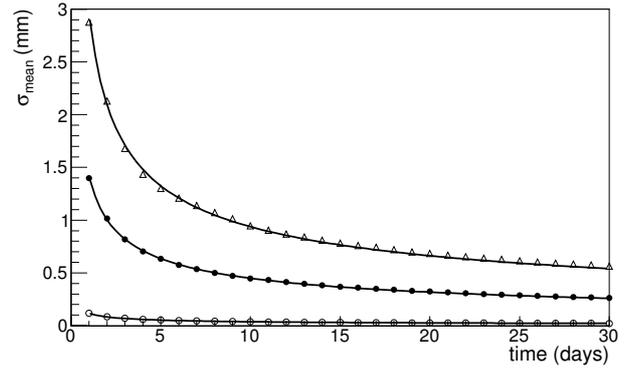


Fig. 4. Relation between the standard uncertainty on the mean value of the sample distribution and the data taking time, for the “muon target” in position P1 (\circ), P2 (\bullet) and P3 (\triangle). Each plot is fitted with the function (4).

sition of the “muon target” relative to the “muon telescope”.

The uncertainty on the mean value is given by the well known relation:

$$\sigma_{mean} = \sigma_{distr} / \sqrt{N_{ev}} \quad (3)$$

where N_{ev} is the number of events in the distribution. Since in the same geometrical conditions the number of events in the sample distribution is proportional to the data taking time, the measurement standard uncertainty depends only on the inverse of the square root of the data taking time. In fig. 4 the relation of the position measurement standard uncertainty and the data taking time for the three examined conditions is plotted up to a data taking time of one month.

As time increases, the measurement standard uncertainty decreases. By fitting the plots with the general relation:

$$\sigma_{mean} = C / \sqrt{t} \quad (4)$$

where C is a constant depending on the geometry and materials interposed and t is the data taking time expressed in days, the following values for the constant C are obtained in the three conditions considered: $0.12 \text{ mm day}^{1/2}$, $1.42 \text{ mm day}^{1/2}$, $2.96 \text{ mm day}^{1/2}$; the errors from the fit procedure are below 1%.

For example, in one month of data taking in position P3, where “muon target” and “muon telescope” are positioned 13.0 m far apart, a measurement standard uncertainty of the order of 0.5 mm may be achieved. The same standard uncertainty may be achieved in a week of data taking in position P2, whereas a 0.1 mm may be measured in just one day in the position P1.

As expected, the standard uncertainty of the measurement system depends on the geometrical configuration considered, since both the root mean square of the distributions and the rate of useful events collected are strongly dependent on the geometry of the system and on the amount of materials interposed. Nevertheless, although requesting different data taking times, the position monitoring of all the three inspected points

by a cosmic ray tracking system could provide performances compatible with the requested precisions and with the time scale characteristic of the inspected deformation phenomenon. Typical time scales, in the case of “Palazzo della Loggia” and, in general, for historical buildings, may span over several years. Furthermore, as demonstrated in this case and unlike other monitoring systems, a stability monitoring system based on tracking of cosmic ray muons can efficiently operate also when the building points to be monitored are not reciprocally visible and are separated by solid masonry structures.

B. Measurement of seasonal deformations of the wooden vaulted roof of the “Palazzo della Loggia”

Due to the low cosmic ray rate, a monitoring system based on cosmic ray muon tracking can't provide high precision results in short time. Therefore, it can be competitive with other monitoring techniques only when the deformation under study develops over periods of months or years and the requirements for the monitoring system is to track the slow deformation with time.

This is the case of the cyclic seasonal deformations of the wooden vaulted roof of the “Palazzo della Loggia”, which have been simulated with the Monte Carlo program for points P1, P2 and P3 of the roof structure. As a realistic model of the seasonal deformation, the measured displacement in point P2 on the arch reins, reported by the measurement campaign [20], [21], was adopted.

In fig. 5 the curve corresponding to the assumed seasonal deformation (the same for the three points) is shown as a continuous line. The results of the simulated measurements of the position of the “muon target”, displaced following the assumed structure deformation, are shown; sampling rates of one week, two weeks and one month respectively for points P1, P2 and P3 have been used. It is evident the ability of the proposed measurement system to follow seasonal structural displacements of few millimeters and, consequently, also systematic ones.

CONCLUSIONS

Cosmic ray muon detection techniques for stability monitoring of historical buildings have to deal with the low rate of muon events and with the stochastic nature of the deviations of the muon trajectories due to multiple scattering in crossing materials.

However, due to the very slow evolution of the deformation phenomena that may characterize the behavior of historical building structures, as in the case illustrated in the present work, these constraints do not really constitute a severe limitation for the employment of the proposed method.

Conversely, the ability of muons to penetrate large thicknesses of material suffering only small deviations of the trajectories offers a new possibility to perform the stability monitoring of parts of the building physically and optically separated by solid structures, as walls or

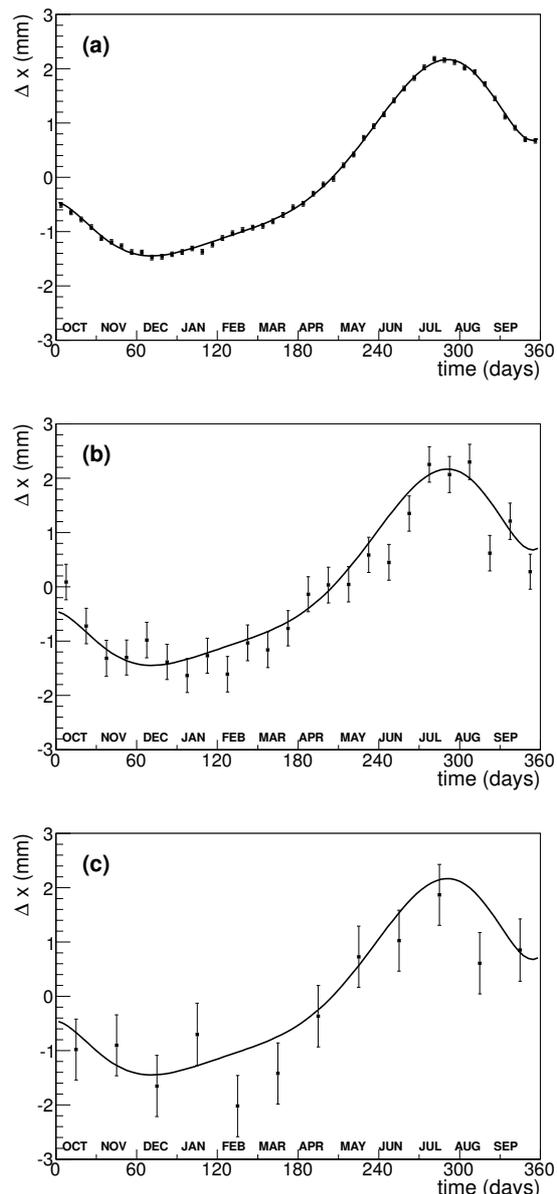


Fig. 5. The seasonal deformation evaluated by the measurement campaign [20], with superimposed the result of the simulation measurements with the sample mean of the position of the “muon target” displaced following the assumed structure deformation, with sampling rate of one week for position P1 (a), two weeks for position P2 (b) and one month for position P3 (c).

floors.

For particular applications, these performances may be competitive in respect to the ones of monitoring systems today widely employed as laser scanner and theodolites, which make use of visible light, or as global position system based methods, hardly applicable to monitor with high resolution internal parts of the building. In addition, whereas the performances of monitoring techniques based on pendulums, inclinometers and extensometers provide measurements of deformation or strain in specific point positions, a global and simultaneous muon monitoring system may be constituted exploiting compact muon detectors distributed in different positions inside the building.

Acknowledgement

This work was supported by a special project of the Department of Mechanical and Industrial Engineering of the Brescia University.

The authors are grateful to G. Baronio of the Department of Mechanical and Industrial Engineering of the Brescia University for his help in the preparation of the present work.

Special thanks go to Prof. Ezio Giuriani and Prof. Alessandra Marini of the Department of Civil, Architectural, Land and Environmental Engineering and Mathematics of the Brescia University for the information provided on the long-lasting study performed on the “Palazzo della Loggia” and for the invaluable advice concerning the problem of historical building monitoring.

REFERENCES

- [1] J. Beringer *et al.* (Particle Data Group), The Review of Particle Physics, Phys. Rev. D 86 (2012) 010001.
- [2] E. P. Georges, Commonwealth Engineer (1955) 455.
- [3] L. W. Alvarez *et al.*, Science 167 (1970) 832.
- [4] H. Tanaka *et al.*, Nucl. Instr. Meth. A 507 (2003) 657.
- [5] K. Borozdin *et al.*, Nature 422 (2003) 277.
- [6] P. M. Jenneson, Nucl. Instr. Meth. A 525 (2004) 346.
- [7] I. Bodini, G. Bonomi, D. Cambiaghi, A. Magalini and A. Zenoni, Meas. Sci. Technol. 18 (2007) 3537.
- [8] D. Gibert *et al.*, Earth Planets Space 62 (2010) 153.
- [9] K. Borozdin *et al.*, Phys. Rev. Lett. 109 (2012) 152501.
- [10] Mu-Steel Project, Project carried out with a grant of the European Commission within the Research Fund for Coal and Steel, RFSR-CT-2010-00033.
- [11] Mu-Blast Project, Proposal submitted to European Commission within the Research Fund for Coal and Steel, 630643.
- [12] D. E. Groom, N. V. Mokhov and S. Striganov, “Muon stopping power and range”, Atomic Data and Nuclear Data Tables, Vol. 76, No. 2, July 2001.
- [13] G. Z. Moliere, Z. Naturforsch. 2a (1947) 133; Z. Naturforsch. 3a (1948) 78.
- [14] H. A. Bethe, Phys. Rev. 89 (1953) 1256.
- [15] S. Agostinelli *et al.*, Nucl. Instr. Meth. A 506 (2003) 250-303.
- [16] R. Brun and F. Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP’96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. Meth. in Phys. Res. A 389 (1997) 81-86.
- [17] <http://root.cern.ch/drupal/content/geant4-vmc>.
- [18] A. Donzella, Nuovo Cimento, *article in press*.
- [19] I. Bodini *et al.* Historical building stability monitoring by means of a cosmic ray tracking system, arXiv:1403.1709 (2014).
- [20] A. Franchi, E. Giuriani, A. Gubana, G. Lupo, G. Mezzanotte, P. Ronca and V. Volta, *Per la conservazione del Palazzo della Loggia di Brescia - Parere sulla stabilità strutturale* Centro di studio e ricerca per la conservazione e il recupero dei beni architettonici e ambientali, Dipartimento di Ingegneria Civile, Università di Brescia (Grafo edizioni, Brescia) 1993.
- [21] A. Bellini *et al.* *Il Palazzo della Loggia di Brescia - Indagini e progetti per la conservazione*, Atti del Convegno di studi: “Storia e problemi statici del Palazzo della Loggia di Brescia”, Università degli Studi di Brescia - Facoltà di Ingegneria, ottobre 2000 (Starrylink editrice, Brescia) 2000.
- [22] L. Bonechi, Proceedings of the 29th International Cosmic Ray Conference, (Pune) (2005) 101.

TOWARDS AN EXECUTABLE SOCIOTECHNICAL MODEL FOR PRODUCT DEVELOPMENT AND ENGINEERING SYSTEMS

Axel Hahn and Jürgen Geuter
Business Engineering
University of Oldenburg
26169, Oldenburg, Germany
Email: {hahn,geuter}@wi-ol.de

KEYWORDS

CSM, Systems theory, hierarchical systems, engineering, social systems

ABSTRACT

Individual skills and properties of the human beings involved have made analyzing product development systems difficult. A flexible approach to modeling and simulating these social systems is still missing. Due to their high degree of structure product development processes do lend themselves to simulation and analysis though. In this paper we outline the specific challenges and requirements for creating a complete model of social product development systems. We also propose a new approach for modeling and simulating real-world development systems.

INTRODUCTION

The modeling and simulation of processes involving human actors grows more complex the more freedom the respective actors have. Simple economical models which reduce human being to rational agents optimizing a simple number such as profit or cost are very simple to build and analyze. But while these simple models do help us understand certain social behaviors or systems better, many of the social contexts we want to model and analyze are inherently more complex, not only in the sheer amount of involved objects and properties but also in the way these objects, properties and goals need to be modeled: Human beings can quite comfortably operate based on a significant number of partially contradicting goals and pinning down a person's properties to a number is not always a suitable approach.

Looking at real-world social processes and systems the domain of product development and engineering provides us with a good opportunity to take a big step towards better understanding social behavior and systems: While engineering processes are characterized by standardized methods and processes, the individual developers' skills and properties as well as the way the whole group communicates and collaborates have

a huge impact on the probability of a given project's success [Song et al., 1997].

Standardized processes, qualifications, methods, tools and metrics have led to a growth in productivity and quality within the engineering domain over the last decades. But even with a rising degree of maturity considering processes and methods, a large number of engineering and design projects fail to either meet their required deadlines or stay within the given resource limitations. Many projects fail to reach their main goal at all. [Project Management Solutions, 2011] found that about 37% of ICT projects were in serious danger of failure, [Heeks, 2003] found that in eGovernment projects "35% are total" and "50% are partial failures". The goal of innovation competes with the more traditional project targets (lead time, cost, etc.): Where Innovation "consists in diverging processes that explore new alternatives, values and performance criteria" [Aggeri and Segrestin, 2007] Project development "consists in converging processes built around predefined targets and deadlines" [Aggeri and Segrestin, 2007]; innovation enforces a certain level of uncertainty.

The problem of project failure has mostly been dealt with from a process- or artifact-based angle: Traditional project management has developed methods to estimate effort and time [Boehm et al., 1995]. More traditional, static process models for engineering such as the Waterfall Model [Royce, 1970] or the V-Model [Broy and Rausch, 2005] have been extended (and in some areas even replaced) by more agile models (i.e. SCRUM) to help management making better estimations about project progress and to allow a quicker recovery from design mistakes. Also the design process itself is being rethought and improved, for example by compartmentalizing and formalizing the different modules of the product as for example Contract-Based Design [Sangiovanni-Vincentelli et al., 2012] suggests.

But studies such as [GPM e.V. und PA Consulting Group, 2006] and [Terry, 2002] show that "For the overwhelming majority of the bankrupt projects we studied, there was not a single technological issue to ex-

plain the failure” [DeMarco and Lister, 1999]. The way we model engineering projects lacks an integration of so-called “soft factors” or what DeMarco and Lister [DeMarco and Lister, 1999] call a project’s “sociology”: The way people interact, exchange ideas and build consensus. Building consensus also clashes with mismatching goals: Where the project as a whole has clear goals regarding quality, effort and cost, each individual might – often for organizational reasons – have significantly different goals working against the overall project’s goals: Where the project targets a very high level of quality to serve as the basis for follow-up projects, the individual worker might be forced to do “just enough” to pass the tests in order to have enough time to fulfill his or her other project’s obligations.

In order to understand the effects described above better and to help engineering projects circumvent or mitigate them we need executable social models for product development that allow estimating the impact of certain properties of any given social engineering system. Some typical characteristics of engineering systems, which we will outline in the next section, support building executable probabilistic models of these specialized social systems.

In the following sections we will analyze the characteristics of modern engineering systems illustrating their aptness for modeling and simulation. We will also propose a strategy to model complex, goal-oriented social systems suitable for simulation.

ENGINEERING SYSTEMS

Engineering or designing a product is, on a very abstract level, the exploration of the design space spanned by the given requirements or objectives.

The space of alternatives (which are possible artifacts fulfilling the given objectives) shrinks while adding or substantiating requirements. Partial product models like design documents or partial implementations can further shrink the design space by creating implicit requirements or dependencies. Apart from purely functional requirements, other objectives (such as the expected time until completion for example) are usually added to the System of objectives as well.

In order to separate the social part of the engineering system from the process- or artifact-focussed part, we adopt the ideas proposed in [Ehrlenspiel, 2009]: The driving force of the development process is the “System of Objectives” (SoO) which is constituted by all given explicit requirements and objectives as well as their logical, formal or technical dependencies and connections.

The SoO is complimented by the “Object system” (OS) containing the partial product models and development artifacts as well as their connections, dependencies and metadata as well as the “Action system” (AS) which consists of the people, tools, resources, processes as well as the actions these entities make. Negele [Negele, 1998] later split the AS into the “Action System” containing the people and resources and the “Process system” (PS) aggregating the different processes within the project.

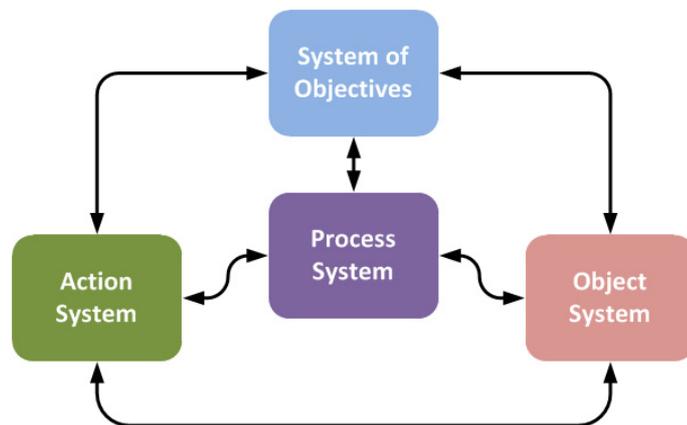


Fig. 1. Breakdown of the product development system into abstract partial systems

All those systems are separate but influence each other as figure 1 illustrates. The different ways that the systems influence each other though is not clearly defined in a general way. The SoO and the OS are tightly coupled with the entities from the OS each fulfilling one or more objectives. Only certain parts of the Action System (people, tools, etc.) can act in any of the PS’s given processes or phases contributing to the completion of one or more goals.

While these very theoretical analyses of the abstract engineering system help structuring further research in the domain, they themselves were not suitable (and intended) to be used to actually model, explain or simulate existing engineering projects.

This systemic approach ignores the sociotechnical aspects of engineering: It is a task heavily based on human skills and usually teamwork. While tools help humans work more effectively and with fewer mistakes, [Song et al., 1997] has shown that “process skills, project management skills, and alignment of skills and needs” have a “strong, positive influence on [...] product performance”. With mastery in a certain area taking many hours of practice and only mastery allowing innovation [Ericsson, 1998] a significant level of specialization has been characteristic of engineering systems: The process of development is often distributed amongst collaborating experts in the area of requirements engineering, design, implementation and testing.

Actual engineering in the wild has not been untouched by the trends of globalization and outsourcing in the last decades. Engineering projects, even those within one company, are increasingly global endeavors, distributing the participants within one project over many different countries. Globalization allows a company not only to build highly specialized development groups while still profiting from potentially lower wages or other environmental factors. But this approach is not without its own, often social, problems. Not only does it require a higher amount of explicit communication due to the lack of the so-called “Watercooler effect” (meaning the kind of communication happening in a company around points of non-work related social interaction, such as a Watercooler) but also opens the

whole process up for cultural mismatches and tensions as exemplified in [Nicholson and Sahay, 2001].

The orientation to project-oriented company organization like the Matrix Organization has put some pressure on the design team as well: With each individual potentially working on not only more than one project but also under more than one leadership, the individual project goals can easily come into conflict with the goals of the different organizational entities or the goals of the individual product designer.

The recent focus on international standards and certifications such as CMMI[Cmami Product Team, 2010] and SPICE (ISO/IEC 15504) have forced globally active enterprises to further solidify and document their methods and processes providing a good foundation for modeling their real-world processes for simulations.

Given the existing research within the area of engineering and the trends we outlined, the domain of engineering and product development provides a very suitable testbed to model and simulate social and sociotechnical processes for the following reasons:

- **Standardized procedures and processes** (Waterfall, Agile development, SCRUM, V-Model, etc)
- **Well-defined process steps** (Component-based development, separation of requirements, design, implementation and testing)
- **Use of standardized metrics** (CMMI/SPICE for degree of maturity, Code Quality Metrics, Error per Line of Code, etc)
- **Division of Labor** (collaborating specialists within each domain such as design, implementation and testing)
- **Distributed development** (globally distributed teams, global collaboration between companies, global component suppliers)
- **Knowledge and skill driven** (importance of individual skills, training of engineers)
- **Social system** (Communication, collaboration)

REQUIREMENTS FOR CREATING EXECUTABLE MODELS OF ENGINEERING SYSTEMS

From sections 1 and 2 we derive the following requirements for modeling engineering systems in a way that allows their execution/simulation corresponding to engineering projects in the wild. We have categorized the requirements in the following subsections: General Requirements for the model, Social Requirements for the modeling of the social (inter)actions and Domain requirements which follow from the domain (engineering/product design) itself.

GENERAL REQUIREMENTS

In general, the model of a sociotechnical model of engineering systems has to consider the following requirements. These are not directly tied to the domain of engineering and can be applied to other similarly structured processes such as for example law-making:

- **Hierarchies of communicating systems:** In order to model the different organizational layers within

complex engineering systems and companies, hierarchies need to be implemented. A system creating a product needs to be able to include the engineering systems building subcomponents. The model needs to be able to implement traditional monolithic design system (one independent R&D department) as well as modern enterprise structures such as Matrix-Organization or development processes spanning many different entities/companies in order to be able to represent actual modern development systems.

- **opaque Systems:** In order to simulate the actions of systems that cannot be made fully transparent (such suppliers of subcomponents or resources) the model needs to be able to deal with opaque, loosely coupled systems. This requirement follows directly from the trend towards globalization as we outlined in section 2. Integrating systems which do not present their internal state or objective allows modeling global distributed engineering without enforcing insurmountable levels of transparency.

- **heterogenous depth:** When modeling a complex engineering system different branches of the hierarchical tree of entities and sub-entities can have - due to the organizational structure of the project or due to a lack of deeper structured information about a part of the project - different depths of sublevels.

- **heterogenous systemic structure:** Each subsystem of a given development project can have its own process and method just as well as its own goals - possibly conflicting with the goals of a higher-ranking system. Each subsystem needs to potentially be observable as its own autonomous unit without sacrificing the opportunity of interconnectedness.

SOCIAL MODEL REQUIREMENTS

When modeling engineering processes it makes sense to make social interactions and factors very explicit. There already are established standards such as SPEM[Object Management Group, 2008] for modeling engineering processes which we need to augment with social influence factors for which we devised the following requirements:

- **Model of communication:** Communication between people and systems does not happen flawlessly. Potential defects or degradation of trust happen due to cultural or language barriers as shown by [Nicholson and Sahay, 2001] or due to the sheer perception of distance [Moon, 1999].

- **Model of knowledge and skills** (as a special form of knowledge): In a domain as dependent on human actions as engineering, a precise model of the acquisition, degradation and application of skills such as [Rasmussen, 1983] and a concept of human memory such as [Shiffrin, 2003] is central for an appropriate simulation.

- **Personal traits:** The model needs to be able to assign different characteristics to human entities within the development system with a focus on modeling biases when it comes to decision making. An example for this is the “perfectionist bias” which highly influences the decision whether to continue to improve a given

artifact or consider its quality high enough. Research [Costa and McCrae, 1992], [Bartle, 2004] shows that reducing the complexity of human behavior for specific tasks to a set of limited archetypes produces good results while providing a simple abstraction for classifying individual human beings.

- **Social interaction and organizational roles:** Product development and engineering happen within social systems and hierarchies. The model needs to take these into account and should be able to model the power structure of a given social system on a structural/organizational level as well as on a social level. This is especially relevant for the propagation of objectives and goals from higher to lower levels within the organizational hierarchy.

- **Other flexible “weak factors”:** Studies [Ernst, 2002] have shown that many non-tangible so-called “weak factors” heavily influence product development success. An executable model of product development needs to be able to potentially include these kinds of influences.

DOMAIN REQUIREMENTS

Finally we developed the following requirements emerging from the domain of engineering itself. Most of these requirements can be fulfilled by adapting or integrating existing standards from the domain itself:

- **Model of artifacts, objectives, goals and their interdependences:** Using an integrated model of product development (as proposed in [Ehrlenspiel, 2009]) which connects the objectives and their fulfillment to the development artifacts is necessary to trace how well a given system follows its set of goals. The model of artifacts also needs to support keeping track of the efforts having gone into the different parts of the product.

- **Model of process:** The structure of processes, their sequence and the potential points where backtracking to earlier steps is necessary needs to be specified. Standards such as for example [Object Management Group, 2008] can provide a solid foundation.

- **Change of requirements:** The executable model needs to be flexible enough to deal with changing requirements or objectives. Legal requirements or a stakeholder’s expectations might change during the run of the project just as a project might need to drop its targeted level of quality in order to meet deadlines. The model needs to keep track of these changing objectives and their consequences for the project itself.

HIERARCHICAL OPAQUE SOCIOTECHNICAL SYSTEMS

Our approach for modeling and finally simulating engineering systems is based on the conceptual work done in systems theory [Luhmann, 2010] that [Ehrlenspiel, 2009] and [Negele, 1998] adapted in an abstract fashion to the domain of product development: The whole engineering system is separated into partial systems (objectives, artifacts, actions, processes, etc.) (see figure 1) which interact through messages and semantic con-

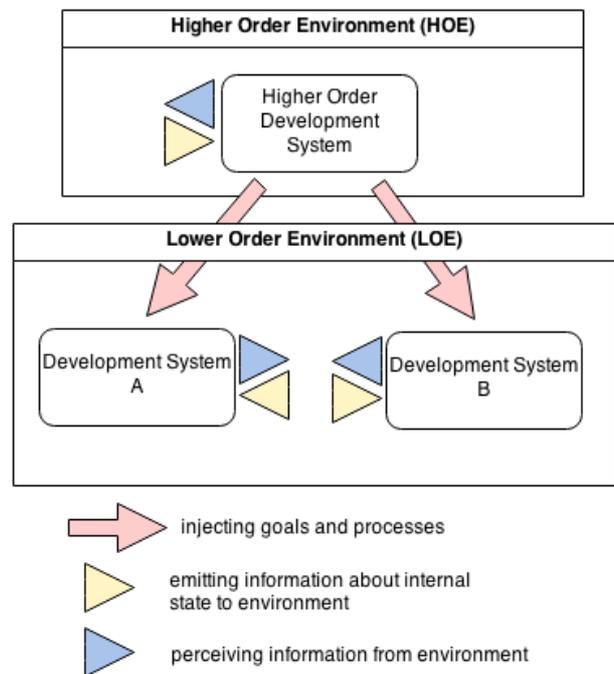


Fig. 2. Illustration of a simple hierarchical model

nections. This allows us to integrate already existing standards and implementations to model parts of the full executable model (like the SPEM standard [Object Management Group, 2008] for process modeling for example).

Using hierarchies of systems and subsystems as an architectural principle is a well established approach, especially for analysing engineering problems. [Varaiya, 1972] developed a theory for hierarchical, multilevel systems which can be applied to cooperating social systems. Finding out how to identify and separate subsystems from each other was further described by [Guinzy, 1973]. Hierarchical systems are a very well-established method for handling systemic complexity within very different systems (Software, Hardware, Social systems, etc).

In our approach we follow the definition of “system” proposed by Luhmann [Luhmann, 2010] in which systems allow no direct access to their innards but only interact with the rest of the world by “irritating” it and each other. This means that each system provides information about itself to the environment and can use other system’s provided data without knowing how another system came to that state. Luhmann’s model of social systems fits in perfectly with Ehrlenspiel’s and Negele’s model while also providing the necessary facilities to model supply chains and distributed development. There are two different ways in which different systems influence each other: Indirectly and directly.

Hierarchical opaque sociotechnical systems (HOSS) (as illustrated in figure 2) consequently connect fully developed singular models to each other either by *direct semantic connections* (such as “is higher-order system to” or “produces sub-part for”) or by emitting information into an environment that is potentially shared with

other development systems. Direct connections allow the propagation of organizational goals and products of the development process whereas shared environments of systems allow the representation of so-called “weak factors” as described by [Terry, 2002]. The systems are opaque meaning that they cannot perceive each other’s internal state or structure but rely on each system communicating its own state through information emitted into the environment or via explicit artifacts (such as reports of achieved level of completion or quality).

The direct way of social systems influencing each other (whether they are individuals communicating or groups of individuals) is based on the idea of social networks: The social systems (either people or groups of people) interact based on a network not only of paths of communication but also based on social and formal hierarchies. Each direct connection between social systems can transport information (potentially based on an artifact as for example a design document being exchanged via email) and each means of transportation of information can have distinctive characteristics such as percentage of misunderstanding or the organizational impact (i.e. a direct order by a superior has a different impact than the suggestion of a coworker).

The hierarchical, systemic approach allows the modeling researcher to model a complex development system in any granularity: For the big picture view only one level of few communicating systems might suffice, for in-depth analyses the hierarchy can be drilled down until the individual human beings are integrated. This also allows combining smaller, well-tested models into bigger, more complex models.

A hierarchical model allows each system to consist of subsystems of a similar type: The Action System “Company” can include multiple Action Systems “Developer” each having their own specific goals and properties next to the ones given to them by the organization. Figure 3 shows an illustration of our approach. Each individual person therefore is their own opaque system with all connected properties and degrees of freedom. Communication between people is modeled by a direct exchange of objects, either artifacts such as a written document or by an abstract “Communication Object” encapsulating a speaking act. The process of a system adapting an object to its own internal state allows the easy integration of models of communication such as [Nicholson and Sahay, 2001] or [Moon, 1999]. Skills and properties are attached to each Action System with each Action System’s environment also containing descriptions of potential changes of properties (as for example a rate of learning/skill acquirement).

Each human within the system is assigned a prototype personality based on typical behavior within product development systems. These prototypes model biases of developers within the development process and are constructed analogously to the way [Costa and McCrae, 1992] and [Bartle, 2004] build their archetypes for other domains by separating individuals based on continuums of different behavioral extremes. In combination with decision heuristics and typical human

non-development-related decision biases as described in [Tversky and Kahneman, 1974] this forms a usable, flexible as well as extensible model of the social and individual aspects of product development.

“Weak factors” that cannot easily be attributed to one entity can either be encapsulated into the system environment or into abstract social systems that emit certain information or state into the social network of communicating and interacting developers. This approach allows the integration of non-tangible effects such as “our company is only 80% effective on Fridays” allowing the analyst of the product development system to explicate certain facts about the system that are known but hard to associate to one source.

The Action system also integrates the use of product development tools: Access to a given artifact within the model can go through a specific tool augmenting a developer’s skills and performance as well as having other properties such as reduced number of errors.

The Systems of Objectives, Artifacts and Processes are not fully opaque but work similar to containers. Each social system knows its own goals and objectives as well as the ones of the social system it is part of. Similarly each social system can have its own artifacts (design document not shared with other groups) and its own processes. Artifacts can be transferred to a different social system through means communication (with the potential of misunderstanding as outlined above).

Development systems with a high degree of maturity will have more information about their inner workings than less mature systems. The approach for modeling development systems as it is described here is extensible enough to include any sort of communication or influence factor without enforcing any.

Because of its structural separation of different systems, the import of existing information from the development tool landscape is possible: Artifact and objective information from a Product Data Management (PDM) system or a source code management system, processes from a process modeling tool or even very simple process models such as those project management tools (MS Project etc.) can provide.

SIMULATION OF HIERARCHICAL OPAQUE SOCIOTECHNICAL SYSTEMS

Right now the HOSS approach does offer a new way to think and structure models of engineering and product development systems. The next step is to create simulators from those descriptive models.

We have implemented simple, prototypical simulators of HOSS models for simple engineering systems with a limited amount of participants, goals, skills and artifacts. Given our experience we are currently in the process of redesigning a more complete and well-defined framework of building blocks for creating HOSS models.

Simulations of HOSS are driven by the hierarchical Action Systems (see figure 3) making decisions according to their skills and their goals concerning the process itself. Applied to the domain the action systems encap-

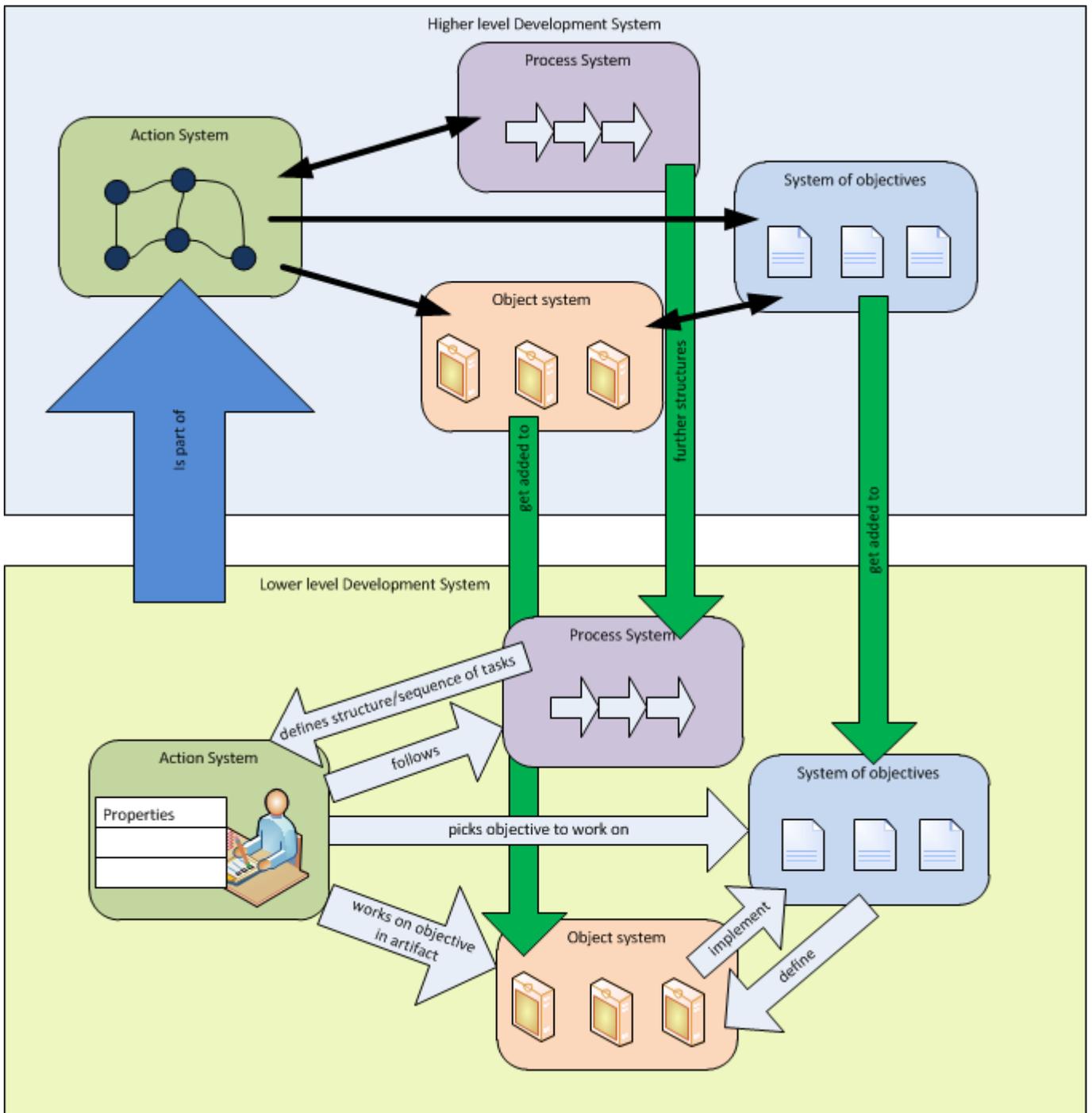


Fig. 3. Illustration of a simple hierarchical model

sulate individual human beings within the development process, groups of people participating in the development process or complete (external) entities such as suppliers, partners, the regulating government or other stakeholders.

The model is based on a discrete understanding of time, with each Action System deciding for each step how to invest their energy in each step of time. While actual work on the implementation of a goal costs a lot of energy, communication is relatively cheap but can create a certain level of stress affecting other properties negatively. Actions can last for more than one step of time depending in the scaling of each step's

duration, this allows integrating systems with a very fine-grained resolution of time with systems who only perform long-lasting actions (such as integrating a local development team whose actions are known within a 10 Minute granularity and an external supplier providing feedback every week).

In order to implement a HOSS model for a real-world development system, it needs to be calibrated to that system. Individual developers or units of developers need to be characterized which can be done either manually or by using techniques like standardized questionnaires. Certain factors of human (inter-)actions can be implemented based on existing psychological and soci-

ological studies and research but have to be evaluated for any given development system. A baseline for how big one unit of work is for any given system has to be derived based on previous projects data as for example outlined in [große Austing and Hahn, 2009].

CONCLUSIONS

The domain of product development and engineering is especially suited for developing a deeper understanding of complex social systems. Its specific characteristics which we outlined in the second section of this paper provide a good compromise of a high degree of individual freedom in decision making as well as a strongly structured and well-defined environment.

Based on existing standards and abstract models hierarchical opaque sociotechnical systems (HOSS) do allow to very precisely model given complex product development systems, thereby helping in analyzing and communicating the structures and dependencies within complex sociotechnical systems.

The structure of HOSS allows the simple implementation of simulators for these complex systems by integrating existing simulations, standard models and statistical models. The next step in our research is to create a set of well-defined building blocks for implementing these simulators as well as a framework to execute these which is the current focus of our work.

The concept of HOSS is not tied to product development but can also easily be adapted to other domains with a similar hierarchical structure such as law-making or other processes of structured decision-making.

HOSS can provide a flexible and simple way to describe and explain social interactions within highly structured environments such as the product development domain.

In order to evaluate the concept more than just theoretically, defining, describing and implementing a simple basic frameworks to implement HOSS is necessary which we already started work on.

REFERENCES

- [Aggeri and Segrestin, 2007] Aggeri, F. and Segrestin, B. (2007). Innovation and project development: an impossible equation? Lessons from an innovative automobile project development. *R&D Management*, 37(1):37–47.
- [Bartle, 2004] Bartle, R. (2004). *Designing Virtual Worlds*, volume p.
- [Boehm et al., 1995] Boehm, B., Clark, B., Horowitz, E., Westland, C., Madachy, R., and Selby, R. (1995). Cost models for future software life cycle processes: COCOMO 2.0. *Annals of Software Engineering*, 1(1):57–94.
- [Broy and Rausch, 2005] Broy, M. and Rausch, A. (2005). Das neue V-Modell® XT.
- [Cmimi Product Team, 2010] Cmimi Product Team (2010). CMMI® for Services, Version 1.3 CMMI-SVC, V1.3 Improving processes for providing better services. Technical Report November, Carnegie Mellon University.
- [Costa and McCrae, 1992] Costa, P. T. J. and McCrae, R. R. (1992). *NEO-PI-R professional manual: Revised NEO personality and NEO Five-Factor Inventory (NEO-FFI)*, volume 4.
- [DeMarco and Lister, 1999] DeMarco, T. and Lister, T. (1999). *Peopleware: Productive Projects and Teams 2nd Ed.* Dorset House Publishing Co., Inc.
- [Ehrlenspiel, 2009] Ehrlenspiel, K. (2009). *Integrierte Produktentwicklung: Denkabläufe, Methodeneinsatz, Zusammenarbeit.* Hanser Verlag, 3., aktual edition.
- [Ericsson, 1998] Ericsson, K. A. (1998). The Scientific Study of Expert Levels of Performance: general implications for optimal learning and creativity. *High Ability Studies*, 9:75–100.
- [Ernst, 2002] Ernst, H. (2002). Success Factors of New Product Development: A Review of the Empirical Literature. *International Journal of Management Reviews*, 4:1–40.
- [GPM e.V. und PA Consulting Group, 2006] GPM e.V. und PA Consulting Group (2006). Ergebnisse der projektmanagement studie konsequente berücksichtigung weicher faktoren.
- [große Austing and Hahn, 2009] große Austing, S. and Hahn, A. (2009). Measurement of product model complexity based on the integrated PLM model. In *Proceedings of The 6th International Product Lifecycle Management Conference*, page 100. University of Bath.
- [Guinzy, 1973] Guinzy, N. (1973). System identification in large scale systems with hierarchical structures. *Computers & Electrical Engineering*, 1:23–42.
- [Heeks, 2003] Heeks, R. (2003). Most e-government-for-development projects fail how can risks be reduced?
- [Luhmann, 2010] Luhmann, N. (2010). *Introduction to systems theory.*
- [Moon, 1999] Moon, Y. (1999). The effects of physical distance and response latency on persuasion in computer-mediated communication and human-computer communication.
- [Negele, 1998] Negele, H. (1998). *Systemtechnische Methodik zur ganzheitlichen Modellierung am Beispiel der integrierten Produktentwicklung.* Utz.
- [Nicholson and Sahay, 2001] Nicholson, B. and Sahay, S. (2001). Some political and cultural issues in the globalisation of software development: case experience from Britain and India. *Information and Organization*, 11(1):25–43.
- [Object Management Group, 2008] Object Management Group (2008). Software & Systems Process Engineering Meta-Model Specification. *Process Engineering*, (April).
- [Project Management Solutions, 2011] Project Management Solutions (2011). Strategies for project recovery.
- [Rasmussen, 1983] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE TRANS. SYS. MAN CYBER.*, 13(3):257–266.
- [Royce, 1970] Royce, W. (1970). Managing the development of large software systems. *proceedings of IEEE WESCON*, 26:1–9.
- [Sangiovanni-Vincentelli et al., 2012] Sangiovanni-Vincentelli, A., Damm, W., and Passerone, R. (2012). Taming Dr. Frankenstein: Contract-Based Design for Cyber-physical Systems. *European Journal of Control*.
- [Shiffrin, 2003] Shiffrin, R. (2003). Modeling memory and perception. *Cognitive Science*, 27(3):341–378.
- [Song et al., 1997] Song, X. M., Souder, W. E., and Dyer, B. (1997). A causal model of the impact of skills, synergy, and design sensitivity on new product performance. *Journal of Product Innovation Management*, 14:88–101.
- [Terry, 2002] Terry, C.-D. (2002). The “real” success factors on projects. *International Journal of Project Management*, 20(3):185–190.
- [Tversky and Kahneman, 1974] Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, 185:1124–1131.
- [Varaiya, 1972] Varaiya, P. (1972). Theory of hierarchical, multilevel systems. *IEEE Transactions on Automatic Control*, 17.

**Simulation,
Experimental Science
and
Engineering in
Maritime Operations**

A HARDWARE-IN-THE-LOOP SIMULATOR FOR OFFSHORE MACHINERY CONTROL SYSTEM TESTING

Johnny Aarseth, Alf Helge Lien and Øyvind Bunes
Deck Machinery Seismic & Subsea
Rolls-Royce Marine
PO Box 193, N-6069 Hareid, Norway
Email: Johnny.Aarseth@Rolls-Royce.com
Alf.Helge.Lien@Rolls-Royce.com
Oyvind.Bunes@Rolls-Royce.com

Yingguang Chu and Vilmar Æsøy
Department of Maritime Technology and Operations
Aalesund University College
PO Box 1517, N-6025 Ålesund, Norway
Email: yich@hials.no
ve@hials.no

KEYWORDS

Offshore machinery, hardware-in-the-loop, modelling and simulation, Bond Graph.

ABSTRACT

The paper presents a concept to develop a simulator for testing of the control systems of automated handling systems for offshore vessels. The concept is based on using Bond Graph for numeric modelling of physical systems, then implementing the model on a platform which can run and visualize simulations and communicate with the target control system in real time. A simulator is implemented and tested for a Launch and Recovery System (LARS) from Rolls-Royce Marine. The intention for this simulator is to provide a platform for development and testing of the control system without having a full scale physical system available, as that can be both practically difficult and an expensive way of performing testing. The simulator can also serve as a platform operational training as well as testing and monitoring the performance of the machinery system effectively, in particularly the characteristics of the hydraulic components during runtime.

INTRODUCTION

Offshore operations involving complex and advanced machinery are never easy to perform due to the challenging and complex interacting environment. Norway is one of the leading players in the world in maritime industry. The rough conditions of the Norwegian Sea area forced the development of many specific task-oriented and innovative offshore solutions, such as anchor handling, pipe laying, deep sea drilling, launch and recovery of remotely operated tools and vehicles, etc. A great deal of resources and efforts are put into investigations, R&D, testing, training and maintenance of offshore machinery systems. For companies developing and supplying such equipment and services to the offshore industry, the degree of success is highly dependent on the reliability of the delivered systems in operation. This comprises both operability and safety functions, and hence the control systems need to go through extensive testing before being installed in new or existing systems. However, the possibility of testing a system in full scale at the work

site is often both limited and expensive. Thus, a lot of tests have to be performed on-board a vessel or a rig, which in such cases limiting the scope of work and time used for testing is important since the delivery time to market is another key factor of a successful product.

Hardware-in-the-loop (HIL) simulation is a form of real time simulation that used for developing and testing of complex embedded systems. Since the last two decades, HIL simulation has becoming increasingly prevalent in many industries driven by the complexity of advanced systems and the costs of both time and capital in building and testing on real systems. The general concept of HIL simulation is depicted in Figure 1. The physical system is replaced by a simulator by reproducing its behavioral characteristic and response to external commands by the use of a mathematic model. During testing the simulator, the target control system should not experience significant difference from being connected to the real system (DNV 2011). For systems that are intended to be visually observed, visualization of the simulator should also be included. With a 3D visualization window the simulator can also be used for operator training purpose, which is an indispensable part during product development in offshore industry as sea testing is both time consuming and expensive to carry out.

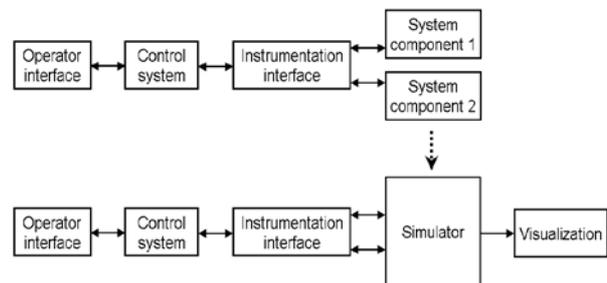


Figure 1: Concept of physical testing & HIL testing

HIL modeling methods and platforms vary from case to case depending on the physical systems and their control systems. Some examples of HIL simulator can be found for different applications including Hydro Power Plant Governor (Zaev, E 2012), Sequential Turbocharging System (Jianwei Du 2007), Engine Electronic Control

Unit (Cebi, A 2005). The project presented in this paper was carried out at the Aalesund University College in close cooperation with Rolls-Royce Marine. The objective is to develop a simulator for testing software and hardware in the control system for the Launch and Recovery System (LARS) as shown in Figure 2. The LARS is used on vessels during subsea operations for handling ROVs from deck to working depth. The system mainly consists of a power unit, a winch and an overboarding unit, which is represented by a traditional A-frame here.

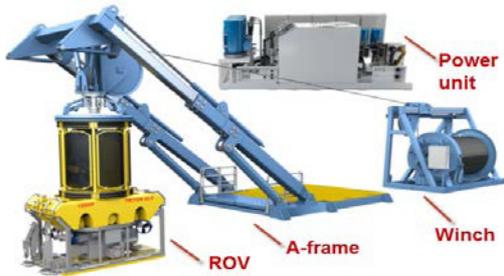


Figure 2: Rolls-Royce LARS for ROVs

The main contributions of this project work lie in two aspects. First, the modelling method based on Bond Graph (E. Pedersen 2008) can be used for modelling of complex systems covering multiple domains, including mechanical, electric, hydraulic and thermal domains. The 20-sim software provides several toolboxes for system modelling and simulation. 20-sim 4C is used for generating C code to run on the real time simulator. Second, the chosen Bachmann M1 controller provides commonly used interfaces for communication with the control systems. Thus, the original control system of the target machinery can be kept identical for the simulator and the real physical system. For most of other HIL simulators found in existing literature, the numeric models and modelling program use block diagrams built in Matlab. Compared to Matlab, 20-sim is a more open software, which allows users to build customized models depending on their needs. Previous researches in multi-domain system modelling and simulation using 20-sim based on BGM can be found in thermal-hydraulic system (Aridhi, E. et al. 2013), electro-mechanical system (Batlle, C. et al. 2008), and mechatronics system (Kayani, S.A. et al. 2007), etc.

The rest of the paper is organized as follows: the architecture of the simulator is described firstly; then the modelling and simulator platform setup is explained; and last, the results of several simulation and testing of the LARS simulator are presented.

LARS HIL SIMULATOR ARCHITECTURE

An embedded hardware solution running the simulator and control system as a stand-alone unit is built to test if the proposed concept would work as intended. The simulator architecture is depicted as in Figure 3. A

numeric model of the LARS is developed and implemented on the Bachmann M1 controller for real time simulation. The simulator is then interfaced with the existing control system of the LARS to receive commands and provide feedbacks representing the response of the LARS.

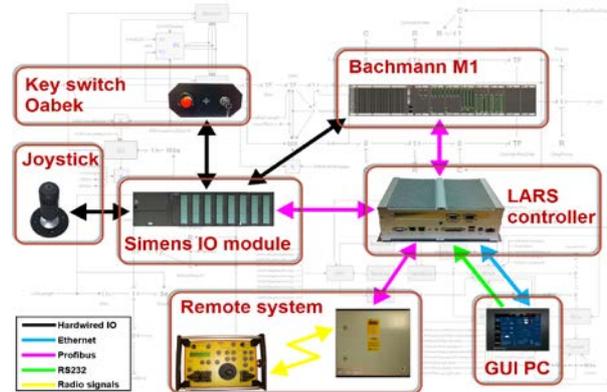


Figure 3: LARS HIL simulator architecture & communication topology

1. Hardware subsystems

To keep the interfaces between the control system and simulator as close as possible to the real system, all interfacing components used in the existing control system of the LARS are kept untouched (personal communication). This is done to maximize the validity of the results when using the simulator for control system testing. To reduce the physical size of the control system hardware, some signals normally transmitted by means of conventional IO were transferred to be transmitted over ProfiBus via Simens IO module. By this approach, the differences between the control system software used for the real system and the simulator came down to mapping of signals between conventional IO and the ProfiBus interface. As both mentioned technologies for signal transmitting were already used in the existing control system, this provided for a flexible solution for testing, as which signals are transmitted through the different communication methods could easily be changed.

For communicating with an external control system the model must run on a real time operating system on a device equipped with the necessary interfaces. Bachmann M1 is a series of industrial controllers supported by 20-sim 4C offering a broad selection of CPU and IO modules, including conventional electrical IO and ProfiBus which are required for the simulator of the LARS HIL project.

The simulator is operated from the same two operator stations as the original control system: an operation panel and a radio remote system. The operation panel consists of a GUI PC and a prefabricated plate with emergency stop, alarm buzzer, key switch, and a

joystick. The graphical user interface used for the simulator is identical to what is used on the real LARS. The only difference between the GUI PC used for the simulator and the real LARS is that the GUI PC for the simulator also runs a Motion Reference Unit (MRU) simulator. The MRU simulator, mimicking the measurement of vessel motion data in 6 DOF, runs as a separate software application not affecting the system's graphical user interface. It is connected to the main controller using RS232 interface. The measured data of the ship motion is used by the control system when operating in active heave compensation mode.

The simulator communication topology and interface types between the hardware subsystems are illustrated as in Figure 3, including Conventional IO (black line), Ethernet (blue line), Profibus (pink line), RS232 (green line) and Radio signal (yellow line).

2. Software tools

The numeric model is built using Bond Graph method in the 20-sim software provided by the Controllab group. The 20-sim software includes a 3D animation feature which is used to build up a graphical model of the LARS for visualization based on shell models exported from 3D CAD software. 20-sim 4C is used to link data from the numeric model to the graphical model during runtime, providing for real time motion of the latter during simulation. It was also used to export the numeric model from 20-sim to the Bachmann M1 controller and to set up mapping between the model variables and the simulator IO interfaces. When using the Bachmann controller in an embedded solution, 20-sim 4C handles connections to the target, mapping of IOs, compilation of target specific code and logging and monitoring of system performance. It has the capability of displaying 3D animations created in 20-sim during runtime by fetching the required variables from the target when running the model.

NUMERIC MODEL

The Bond Graph Method (BGM) is a modelling method to graphically describe the dynamics of a physical system by identifying the energy flow through the system. It can be used for modelling systems across different energy domains including the mechanical, electrical, hydraulic and thermal domains. The method uses a set of ideal elements to represent separate physical phenomenon's which can be combined to make models of complex systems. As the calculation power required in solving the equations of a model increases with its complexity, the level of detail must be adapted appropriately to the model's purpose to achieve a real time simulation without compromising the real world behaviour.

1. Modelling of the LARS

The Bond Graph model for the LARS is built up in main modules similar to its physical architecture including a

hydraulic power unit, winch and A-frame, as illustrated in Figure 1. Furthermore, these modules consist of sub-modules at varying levels, depending on the level of detail in which the component is modelled. The model includes the hydraulic and mechanical dynamics of the system, and in addition an interface to the control system is implemented. The hydraulic submodels are created based on the basic principles of fluid dynamics (ASSOFLUID 2007).

For the mechanical domain, models representing a ROV and the steel armoured umbilical connecting it to the winch were also created. The requirement for the model as regards to interfacing is defined by the interface of the control system software. As the size and complexity of the model sets, the requirement for processing power which affects the real-time performance, the level of detail for different parts of the model have been simplified to different extents.

The model is grouped in modules arranged similarly to the physical main components of the system, as illustrated in Figure 4.

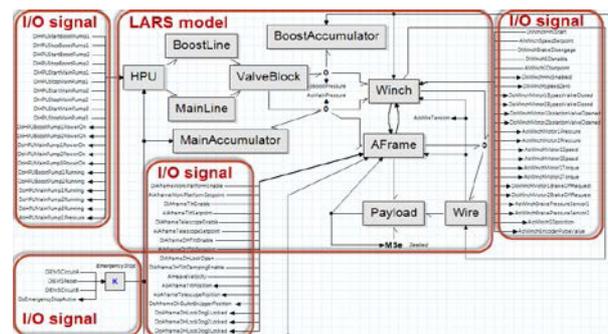


Figure 4: Bond Graph model overview

The HPU module provides hydraulic power to the system and feeds oil through pipe lines to a valve block, where the oil is distributed to the winch and A-frame. Accumulators are connected to the hydraulic system to cover rapid flow peaks, but are too small to store and release significant energy during active heave compensation. The winch and the payload are connected through an umbilical, and hence the winch handles the payload when not resting on other physical objects such as the seabed or vessel deck. The A-frame is used for handling the payload when docked in to the docking head lock, through which the umbilical between the winch and payload is routed. Position of the A-frame thereby affects the umbilical routing, contributing to handling the payload also when the latter is undocked. As a result, the A-frame has power bonds connected to both umbilical and payload. Vessel deck and seabed are included as vertical supports connected directly to the payload and releasing tension from the umbilical upon contact. Vessel motion affects the positioning of the system and hence the umbilical outlet and is included in the A-frame. All signals listed are included to support

the interface required by the control system. A signal based emergency stop circuit is connected to the main components activating certain safety functions.

For each submodel (block) of the LARS, the components are modelled as bond graph or equation submodel describing the energy flow through the system. Details of the submodels are not shown due to the limitation of the paper size.

2. 3D visualization

For visual presentation of the LARS, a built-in 3D animation toolbox is provided in 20-sim. 3D animation toolbox is a 3D visualization tool which can be used to create simple three dimensional figures or import detailed shell models in the STL format, which is supported by most 3D CAD tools. The 3D animation concept is built up using reference frames to achieve the movements of physical parts in relations with each other in a hierarchy nestled to a common base reference frame. All imported and created parts used for visualization has a dedicated reference frame as illustrated in Figure 5. Each frame can be assigned with fixed or variable positioning and alignment to its parent frame. Variable positioning or alignment can be linked to any parameter or variable in the numeric model of the simulator during runtime.

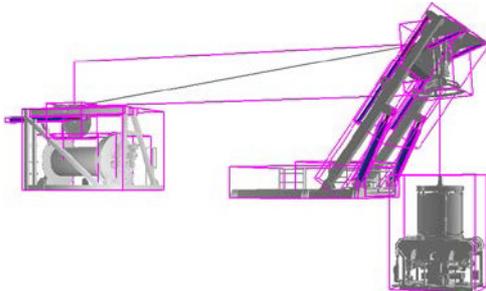


Figure 5: LARS reference frame setup

SIMULATOR SETUP

As the setup is interfaced with the original control system of the LARS, it provides for flexibility as to physical realization of the interfaces and can vary from including just a few of the main parts communicating on ProfiBus and Ethernet to a full scale control system with all IOs connected. The components are fit into a compact cabinet with joystick and GUI PC mounted on the front door (Figure 6). For testing modifications of the LARS as to mechanical and hydraulic design and how this affects the system including the control system, a simulator in its simplest form with only the least required interfacing is needed. The same goes for testing the control system software, unless testing is related to handling of errors in the electrical hardware. For testing electrical hardware the physical interfaces can be set up to include the necessary components.

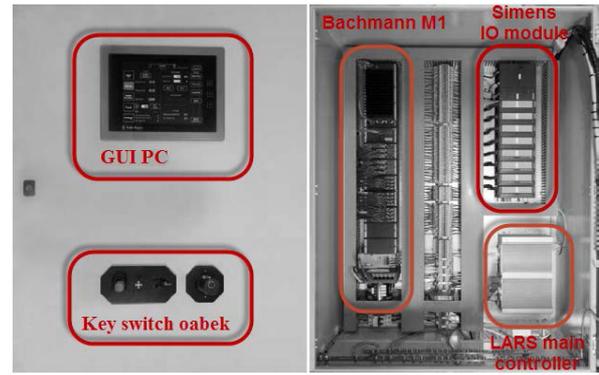


Figure 6: Simulator set up.

MODEL BEHAVIOR AND SIMULATION

To verify that the simulator performed as intended, various tests related to behavior and performance were executed. The main task for the LARS is handling ROV between vessel deck and sea surface. During launch and recovery when the ROV is locked into the docking head, the system is normally in Latched mode. This is a mode where the control system runs the winch based on the tension measurement in the umbilical to maintain a constant tension of approximately $2T_e$. This allows for the operator to manouvre the A-frame without considering the winch simultaneously to avoid slack or over-tensioning the umbilical. The umbilical tension has to be below $2T_e$ to be allowed to set the winch to Latched mode. The tests presented in this chapter were performed using full hydraulic power with all pumps running, while the tilt cylinders of the A-frame were used to move the A-frame in and out.

Prior to the test the A-frame was placed over the vessel side in overboarding position 4, which means 4 m from zero position in positive direction. The overboarding zero position is defined as the position of the docking head lock when the A-frame telescope structure and the docking head is aligned at 90 degrees vertical. Towards the vessel centre is defined as negative direction while over the vessel side is defined as positive direction as illustrated in Figure 7.

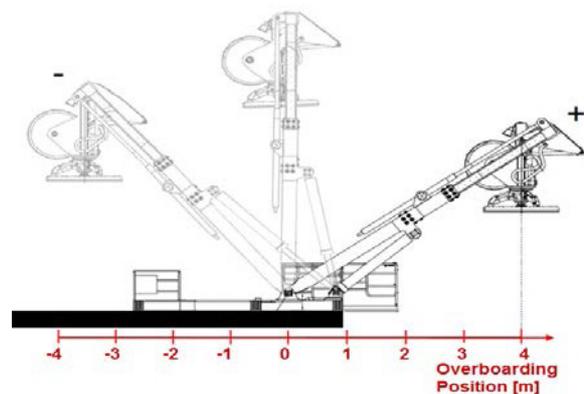


Figure 7: A-frame docking head overboarding position

The test started with the winch in normal mode, which means its speed is controlled by a joystick. Next, the winch was run inwards until the payload was latched into the docking head lock. Once latched, the winch was run out in order to transfer the weight from the winch (umbilical) to the docking head lock (A-frame). When the winch loads drops below $2 T_e$, latched mode was enabled and the A-frame tilt was operated to move inwards until the entire payload was over the vessel deck. The winch was then set back to normal mode and the payload was lifted off the docking head lock. The lock was then opened and the winch was run outwards until the payload was landed on deck. Plots of cylinder pressure, umbilical tension, winch speed and ROV position are shown and explained in next Figure 8.

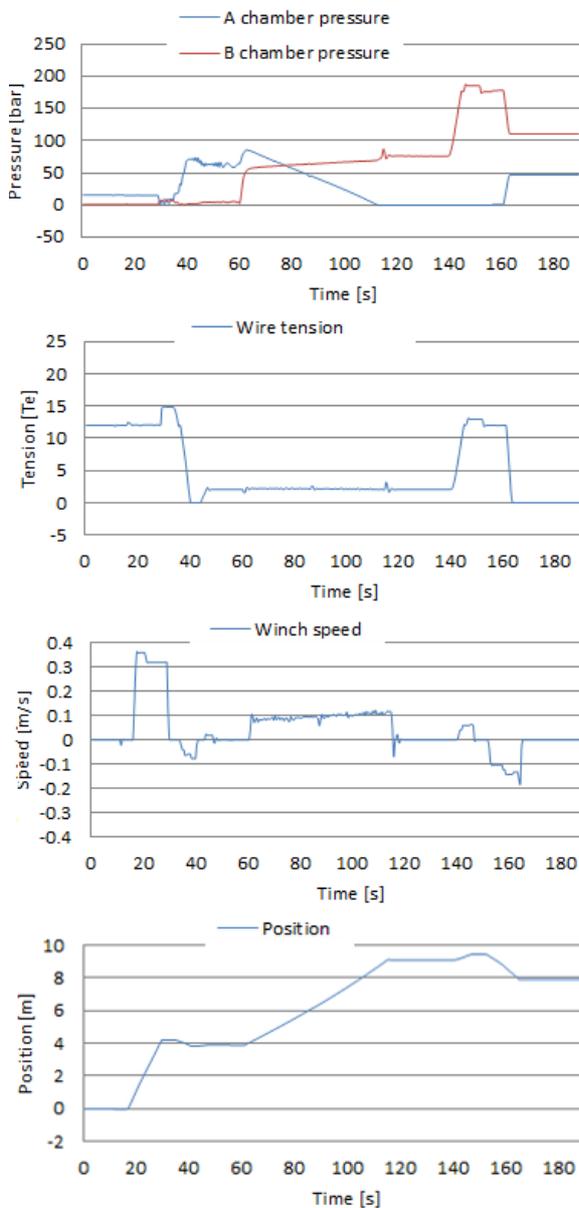


Figure 8: Plotting of cylinder pressure, umbilical tension, winch speed and load position during simulation

Initially, the system was run without any input for about 20 seconds. Then the joystick was used to hoist the ROV as seen in the winch speed and position plots. The ROV was lifted about 4m before it comes into contact with the docking head at approximately 30s. This contact will then try to move the A-frame inwards, thus the tilt cylinders went from being compressed to being stretched as seen in the pressure plot from one of the cylinders. The hydraulic pressure rose in chamber B and fell in chamber A until the docking head lock was closed and the winch lowered the ROV onto the lock. For the tilt cylinder, this leads to the pressure of chamber A settling at a higher level than when only supporting the weight of the A-frame only. The peak in tension around 30s was caused by the ROV coming in contact with the A-frame. The umbilical tension fell rapidly to zero when the ROV was landed on the docking head lock as the weight of the ROV was transferred to the A-frame. When initiating Latched mode, the winch is automatically paying in slowly until the tension in the umbilical reaches approximately $2T_e$ as can be seen in the winch speed and umbilical tension plots. When moving the A-frame inwards, the pressures in both tilt cylinder chambers increased then the A chamber pressure slowly decreased towards zero while the B chamber pressure slowly increased until the A-frame movement stopped at about 120s. The tension plot verifies the control system's constant tension controller was able to maintain an umbilical tension of $2T_e$ while operating the A-frame. From 140 to 150s, the winch was hoisted to lift the ROV off the docking head lock. With the ROV being suspended in the umbilical, the lock was opened and the payload slowly lowered to the vessel deck.

One of the selling points of this LARS is the Active Heave Compensation (AHC) function which expands the operation window in rough weather conditions and thus reduces the potential cost of operation downtime. To test if the model would be able to run in AHC mode within the given range several tests were performed under different wave conditions.

The following pictures show the ROV position during AHC mode with for two different wave motion conditions. In case one (Figure 9), the wave height and period are 2.4m and 5s, respectively. AHC mode started after about 10s and after another 5-10s the heave compensation is fully operational and the ROV position is steady. In case two (Figure 10), the wave height and period are increased to 15m and 10s, respectively. According to the plotting results, this is beyond the capacity of the system and it thus becomes important that the system do not fail, but maintains to compensate within its capability.

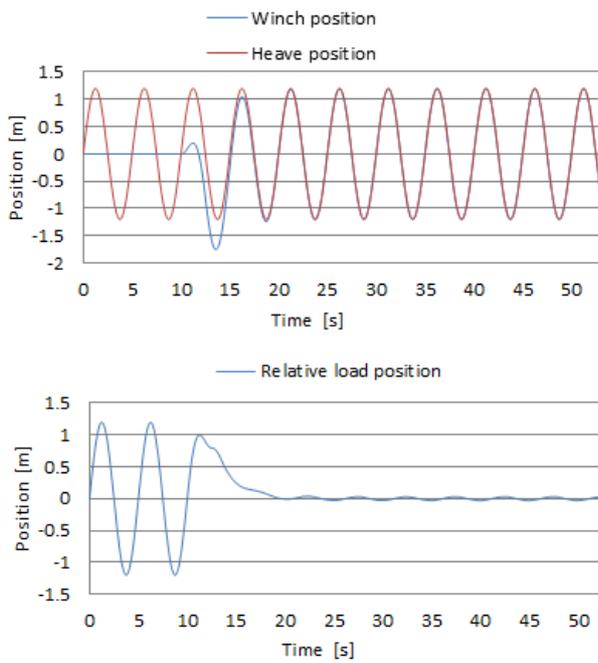


Figure 9: Load position under AHC mode with 2.4m wave height and 5s period

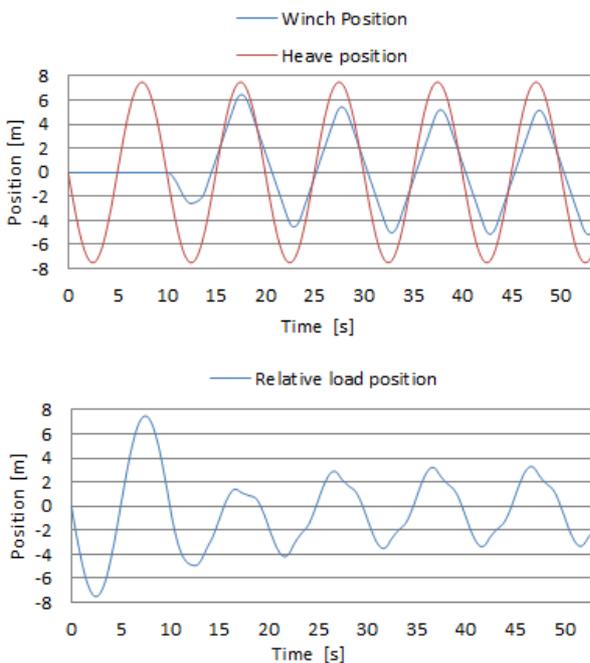


Figure 10: Load position under AHC mode with 15m wave height and 10s period

The control system warned the user about the inability to fully compensate, and the residual motion of the ROV is significant. To verify the control system still works as intended even without being able to compensate the heave, a plot of the winch speed is shown (Figure 11). The speed required for full compensation would be close to 6 m/s, while the actual maximum winch speed is 2.4 m/s.

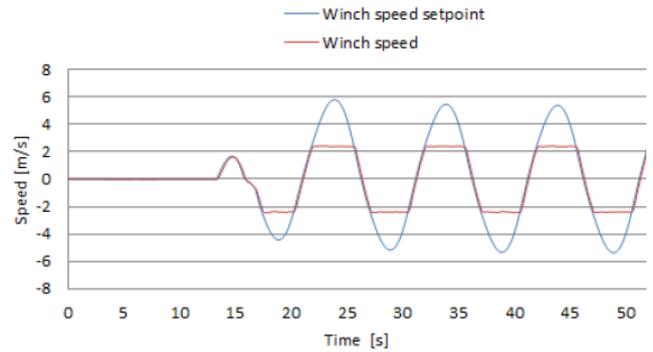


Figure 11: Winch speed under AHC mode with 15m wave height and 10s period

Besides the plots above, all the parameters and variables in the numeric model can be plotted and connected to the 3D model for animation. The model includes the main components of the LARS thus the characteristic response of the system, in particularly the hydraulic performance, can be monitored during and after runtime. Within the field of training, this setup is well suited as it provides for holding courses for troubleshooting as well as operations (Figure 12).



Figure 12: Real time simulation with 3D visualization using joystick, keyboard and radio remote

CONCLUSION

Results from testing the simulator shows the data acquired during simulation runtime corresponds well to expected values related to the performance of a real LARS. This indicates the simulator is well suited for testing of control systems as the response of the system is important when testing control loops and also error handling. The concept developed is modularized similarly to the real system, and utilizes a flexible interface between the simulator and the control system with respect to scalability and interface technology. This means similar simulators can be developed for other systems using the same concept following the same methods as used in this project.

Despite the capabilities of 20-sim as regards to displaying 3D-animation being very good for development, testing and demonstrations, it has some limitations in terms of visual quality. If a simulator should be presented and sold as a commercial product, the visualization should be made more realistic with possibilities to add effects like for instance different

weather scenarios. Especially for an operator training simulator the looks are important to create an environment which feels as close to reality as possible. This can be solved by letting a more advanced graphical motor render the graphics. All parameters, variables and signals in a running model are available for monitoring both in Bachmann Solution Center and 20-sim 4C which communicates with the controller through Ethernet. Hence, setting up an UDP stream from the M1 controller's Ethernet interface with data required by a prospective 3rd party graphical motor can be a way to solve the link to higher quality graphics.

ACKNOWLEDGEMENT

The project was carried out as the authors' thesis assignment of their master study program at Aalesund University College (AAUC) in the cooperation with Rolls-Royce Marine. The authors wish to express their gratitude to the Controllab and Bachmann group for their time and support in performing the usability test. Special thanks go to Prof. Houxiang Zhang from Aalesund University College for his contributions to the paper.

REFERENCES

- Aridhi, E.; Abbes, M.; Mami, A. 2013. "Pseudo bond graph model of a thermo-hydraulic system". In *Proceedings of 2013 5th International Conference on Modeling, Simulation and Applied Optimization* (Hammamet, Apr.28-30), IEEE, Hammamet, 1-5.
- ASSOFLUID, Italian Association of Manufacturing and Trading Companies in Fluid Power Equipment and Components. 2007. *Hydraulics in industrial and mobile applications*.
- Battle, C.; Dòria-Cerezo, A. 2008. "Bond graph models of electromechanical systems. The AC generator case". In *Proceedings of 2008 International Symposium on Industrial Electronics* (Cambridge, Jun. 30-Jul. 02), IEEE, Cambridge, 1064-1069.
- Cebi, A.; Guvenc, L.; Demirci, M.; Karadeniz, C.K.; Kanar, K.; and Guraslan, E. 2005. "A Low Cost, Portable Engine Electronic Control Unit Hardware-in-the-Loop Test System" In *Proceedings of the 2005 International Symposium on Industrial Electronic* (Dubrovnik, Jun.20-23), IEEE, Dubrovnik, 293-298.
- DNV. 2011. *Hardware in the Loop Testing*, Standard for certification No.2.24. DET NORSKE VERIT AS, Høvik. (Jul).
- E. Pedersen; and H. Engja. 2008. "Mathematical Modeling and Simulation of Physical Systems". Lecture notes in course TMR4257 Modeling, Simulation and Analysis of Dynamic Systems, Department of Marine Technology, Norwegian University of Science and Technology. Trondheim. (Aug).
- Jianwei Du; Yinyan Wang; Chuanlei Yang; and Hechun Wang. 2007. "Hardware-in-the-loop Simulation Approach to Testing Controller of Sequential Turbocharging System", In *Proceedings of the 2007 International Conference on Automation and Logistics* (Jinan, Aug. 18-21), IEEE, Jinan, 2426-2431.
- Kayani, S.A.; Malik, M.A. 2007. "Automated Design of Mechatronic Systems using Bond-Graph Modeling and Simulation and Genetic Programming". In *Proceedings of 2007 International Bhurban Conference on Applied Sciences & Technology* (Islamabad, Jan. 8-11), IEEE, 104-110.
- Zaev, E.; Tuneski, A.; Babunski, D.; Trajkovski, L.; Nospal, A.; and Rath, G. 2012. "Hydro power plant governor testing using hardware-in-the-loop simulation" In *Proceedings of the 2012 Mediterranean Conference on Embedded Computing* (Bar, Jun. 19-21), IEEE, Bar, 271-274.

AUTHOR BIOGRAPHIES

Johnny Aarseth studied Product and System Design (M.Sc) at Aalesund University College. He now works as Design Manager at Rolls-Royce Marine AS.

Alf Helge Lien studied Product and System Design (M.Sc) at Aalesund University College. He works as Senior Project Engineer Automation at Rolls-Royce Marine AS.

Øyvind Bunes received his M.Sc in Marine Technology from the Norwegian University of Science and Technology in 1996. Until 2002 he was employed as a researcher at MARINTEK until he joined ODIM, later Rolls-Royce, in 2002. He now serves as Senior Principal Engineer at Rolls-Royce Marine AS. He is also an Associate Fellow in the Rolls-Royce Engineering Fellowship and an Assistant Professor at Aalesund University College.

Yingguang Chu received his master degree in Product and System Design from Aalesund University College in 2013. He continues as a research assistant and Ph.D. candidate at Aalesund University College.

Vilmar Æsøy received his Ph.D. in Mechanical Engineering in 1996 from Norwegian University of Science and Technology. From 1997 to 2002 he worked as researcher in Aker Maritime and R&D manager in Rolls-Royce Marine AS. Since 2002 he works as an Associate Professor at Aalesund University College.

A SHIP MOTION SHORT TERM TIME DOMAIN SIMULATOR AND ITS APPLICATION TO COSTA CONCORDIA EMERGENCY MANOEUVRES JUST BEFORE THE JANUARY 2012 ACCIDENT

Paolo Neri
University of Pisa
Largo L. Lazzarino n.2
56122 Pisa (Italy)
paolo.neri@for.unipi.it

Mario Piccinelli
University of Brescia
Via Branze 38
25123 Brescia (Italy)
mario.piccinelli@gmail.com

Paolo Gubian
University of Brescia
Via Branze 38
25123 Brescia (Italy)
paolo.gubian@unibs.it

Bruno Neri
University of Pisa
Via Caruso 16
56122 Pisa (Italy)
b.neri@iet.unipi.it

KEYWORDS

Simulation model, Ship manoeuvre, Costa Concordia.

ABSTRACT

In this paper we will present a simple but reliable methodology for short term prediction of a cruise ship behaviour during manoeuvres. The methodology is quite general and could be applied to any kind of ship, because it does not require the prior knowledge of any structural or mechanical parameter of the ship. It is based only on the results of manoeuvrability data contained in the Manoeuvring Booklet, which in turn is filled out after sea trials of the ship performed before his delivery to the owner.

We developed this method to support the investigations around the Costa Concordia shipwreck, which happened near the shores of Italy in January 2012. It was then validated against the data recorded in the “black box” of the ship, from which we have been able to extract an entire week of voyage data before the shipwreck. The aim was investigating the possibility of avoiding the impact by performing an evasive manoeuvre (as ordered by the Captain some seconds before the impact, but allegedly misunderstood by the helmsman). The preliminary validation step showed a good matching between simulated and real values (course and heading of the ship) for a time interval of a few minutes.

The fact that the method requires only the results registered in the VDR (Voyage Data Recorder) during sea trial tests, makes it very useful for several applications. Among them, we can cite forensic investigation, the development of components for autopilots, the prediction of the effects of a given manoeuvre in shallow water, the “a posteriori” verification of the correctness of a given manoeuvre and the use in training simulators for ship pilots and masters.

1. INTRODUCTION

The problem of simulating the ship motion under the effect of propellers and external forces (wind and drift are the most important) is a central question in the field of naval engineering and training centres for ship pilots and bridge officers. In (ITTC 1999) two different typologies of ship manoeuvring simulation models are described, together with the guidelines for their validation. The distinction is made between models for prediction of ship manoeuvrability (predictive models) and models for ship simulators (simulator models). The targets are quite different.

In the first case, “...prediction of standard ship manoeuvres is needed at the design stage to ensure that a ship has acceptable manoeuvring behaviour, as defined by the ship owner, IMO (International Maritime Organization) or local authorities”.

In the second one “Simulator, or time-domain, models are used in real-time, man-in-the-loop simulators, or fast-time simulators for training of deck officers or investigation of specific ships operating in specific harbours or channels”.

The first type of model is generated before the ship is built with the aim to foresee, at design stage, some typical manoeuvring characteristics defined by IMO, such as the radius of turning circles or the parameters of Zig-Zag manoeuvres.

The second type of model can be realized after the ship is built up and some mandatory manoeuvrability tests have been performed. This type of model is capable to operate in time domain and can be utilized in simulators for the training of pilots or for other purposes, like that described in this paper.

For simulator models a sort of “black-box” approach can be used in which the set of parameters needed in the equations of motion is directly extracted from the experimental data.

So-called black-box models are widely employed in many application areas. They are built by listing the external and internal factors which could influence the behaviour of the simulated system, by writing a more

or less domain-dependent mathematical model and by setting its coefficients to values determined by means of some fitting process which makes use of experimental data collected on the real system.

The procedures used to determine the coefficients are well established, and in their simplest form date back to the interpolation theory. Different choices of basis functions (such as polynomials, exponentials of Gaussian functions) present varying flexibilities with respect to the ability to describe some system behaviours (Maffezzoni et al., 1995). Recently, such approximation theories as neural network approximation and machine learning approaches in general have demonstrated robust properties in modelling complex phenomena and systems (Poggio et al., 1990). All of them, unfortunately, show a tendency to diverge from the real behaviour when used outside the interval in which the coefficients have been fitted. To this end, integration schemes based on stiffly-stable methods can be of help in delaying the onset of such divergence (Gear et al., 1971).

In the present application, the experimental data used are the results of the manoeuvrability tests performed before the boatyard delivery of the ship to the owner. Several international organizations are qualified for granting the manoeuvrability certificate (the so-called Manoeuvring Booklet); among them, the CETENA S.p.a. is the Italian Company of Fincantieri Group in charge of certifying the compliance of ship sea trials with IMO regulation.

There are a lot of papers, reports, researches available in literature concerning predictive models which are very important for ship designers and naval architects, as well as for boatyards (Revestido et al. 2001; Sutulo et al. 2002; Oltmann 1996; Ishiguro et al. 1996; ITTC 1996, ITTC 2002a, ITTC 2002b). Conversely, there is very little material concerning time domain simulators of ship motion (Gatis et al. 2007; Shyh-Kuang et al. 2008; Mohd et al. 20012, Damitha et al. 2010). Here we will briefly comment two of them.

In (Mohd et al. 2012), after extracting the equations of motion in which the forces and moments acting on hull appear together with those induced by propellers and rudders, the authors perform some numerical simulations. Anyway, the authors do not present any form of validation of the simulator, by means of comparison between simulated and experimental data. In (Damitha et al. 2010) the authors describe a true real time simulator, a software tool that simulates the motion of the ship and could be used, for instance, for pilots and masters training. The approach used by the authors is very similar to that of this work. In fact, as the authors say, "*The simulation system consists of a ship motion prediction system using a few model parameters which can be evaluated by means of standard ship manoeuvring test or determined easily from databases such as Lloyd's register.*" In the paper, the validation is performed by the sea trials of the oil tanker ESSO Osaka (turning circle), but the simulations result in relevant errors: the simulated turning circle radius is about 550 m while the measured value amounts to about 375 m. These results can be

compared with those reported in Figure 2 to give a preliminary idea of the reliability of the simulator presented in this paper.

We can conclude that, as far as the authors know, there is no evidence in literature of a real time simulator with reliable behaviour and an extensive characterization of the errors.

After this introduction, in Section 2 of this paper we will describe the data used to extract the model parameters and to validate the reliability of the simulator. Then, in Section 3 we will describe the motion equations which allowed us to calculate the effect of the forces (internal and external) applied to the hull, whereas in Section 4 we'll present some preliminary results. Finally, Section 5 contains some conclusions together with a description of future improvements of the model and a list of possible applications.

2. MANOUEVRING BOOKLET AND VOYAGE DATA RECORDER (VDR) DATA

In this section a brief description of the experimental data used in this work is given. These data were used to extract the model parameters and to validate the simulator by estimating the entity of the errors between simulated and experimental data.

Manoeuvring Booklet

The Manoeuvring Booklet of a ship reports the results of sea trials performed by the ship before its delivery. Some mandatory tests have to be performed, such as the ones listed in the IMO resolution A - 160 (es IV) of 27/11/1968 and following release. The main tests are:

- 1) "Turning test", used to determine the effectiveness of rudder to produce steady-state turning (Figure 2).
- 2) "Free stop", carried out by suddenly halting a fast moving ship and measuring the space needed to come to a full stop.
- 3) "Crash-stop", similar to the previous one but also including some measures to come to a stop in a shorter time (for example by backing power).
- 4) "Zig-Zag manoeuvre", performed by moving the rudder from central position to an assigned angle (usually 10 or 20 degrees) alternatively to starboard and to port (Figure 3).
- 5) "Williamson turn", the manoeuvre performed to recover a man overboard (Figure 4).

The sea trials were confined to the area covered by the Differential GPS test range in order to obtain accurate tracking information. This way the position error is reduced to a maximum of 2 m.

During the trials, the position, heading (measured with a gyrocompass), rudder angles, propeller revolution per minute (Costa Concordia had two rudders and two propellers) and shaft power were recorded with a sampling frequency of 1 Hz and stored for further elaborations.

VDR data

Voyage Data Recorders (VDRs) are systems installed on modern vessels to preserve details about the ship's

status, and thus provide information to investigators in case of accident. The ongoing data collection is performed by various devices, and the actual recording of this information is entrusted to an industrial grade computer (Piccinelli and Gubian 2013).

The use of VDRs on ships is subjected to the regulations contained in chapter V on "Safety of Navigation" of the "International Convention for the Safety of Life at Sea" (SOLAS). This chapter has been amended in 1999 to adopt the IMO resolution A.861(20) "Performance Standards for Shipborne Voyage Data Recorders (VDRs)" (IMO 1997). These regulations, entered into force on July 1st, 2002, specify the kinds of ships that are required to carry Voyage Data Recorders; all in all, the list encompasses almost any medium-to-bigger sized ships currently at sea.

The IMO resolution also sets requirements about the operation of the VDR and the kind of information it is required to store. Among the mandatory list, most interesting elements for our analysis are surely the position, speed and heading of the ship, the date/time referenced to UTC, the radar data and the ship automation data, such as the speed of the propellers and the position of the rudder(s).

These informations are collected from a large network of different sensing devices scattered all around the ship. The core of the VDR system is the "concentrator", usually an industrial grade computer, which collects all the data and stores it in a digital memory. The concentrator is powered by the ship's electrical system and also sports a dedicated power source which allows it to work at least 2 hours after the loss of main power. At last, a copy of the last 12 hours of recording is also maintained inside the digital memory of the FRM (Final Recording Medium), a rugged capsule designed to survive an accident and thus be recovered by the investigators.

3. MATHEMATICAL MODEL

The aim of the model is to correlate the input given by the pilot (i.e. motors' power and rudders angles) to the output of the system (i.e. ship's position and orientation). The really complex physical system of the ship has been simplified and schematized with a three-Degrees-Of-Freedom (DOF) model which is adequate for our purpose and consists of the short term prediction of the trajectory of the ship (Maimun et al 2011). Another simplifying assumption is that the effect of wind and current are negligible for short-term simulations. However, these effects have been studied and modelled in the past and could be taken into account in a future improved version of the simulator. Figure 1 represents the chosen DOFs (east, north coordinates and heading ϑ) and their meaning. Using the model of Figure 1, z translation and pitch and roll angles are neglected: those DOFs are crucial for ship's stability studies, but are not relevant for positioning estimation.

Starting from this simple model, we could write the three Newton equations that correlate the accelerations

along the three DOFs with the forces acting on the ship:

$$\ddot{\vartheta} = \frac{M_z}{I_\vartheta} = \frac{1}{I_\vartheta} (M_\delta + M_r + M_p) \quad (1)$$

$$\ddot{x} = \frac{F_x}{m} = \frac{1}{m} (F_u + F_p + F_{\vartheta,u}) \quad (2)$$

$$\ddot{y} = \frac{F_y}{m} = \frac{1}{m} (F_v + F_{\vartheta,v}) \quad (3)$$

where M_δ , M_r , M_p represent the torque made by the rudders, the water viscosity and the propellers respectively. F_u , F_v represent the viscous force made by the water in x and y direction respectively. F_p represents the force made by the propellers. $F_{\vartheta,u}$, $F_{\vartheta,v}$ are the apparent forces (e.g. centrifugal force) due to the ship rotational speed ($\dot{\vartheta} = r$). I_ϑ represents the moment of inertia along the z axis and m represents the ship's mass.

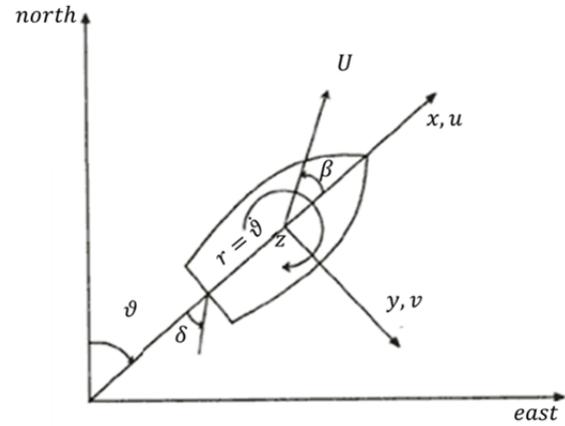


Figure 1: Reference frame.

The exact knowledge of those quantities, and an accurate calculation "a priori", would need a deep study of the ship geometry and of the water conditions (Maimun et al. 2013). Anyway, this work is not aimed to predict the ship behaviour starting from design parameters. Since a sufficient amount of manoeuvring data is available from the Manoeuvring Booklet (CETENA files in the case of Costa Concordia) and from the ship's VDR, it is possible to estimate the relation between the main quantities with an interpolation algorithm.

Eqs. (1), (2) and (3) can then be rewritten in these simple expressions:

$$\ddot{\vartheta} = N_1 \sin(\delta) u|u| + N_2 r|r| + N_3 (p_1|p_1| - p_2|p_2|) \quad (4)$$

$$\ddot{u} = X_1 (p_1|p_1| + p_2|p_2|) + X_2 u|u| + X_3 r^2 + X_4 \ddot{\vartheta} \quad (5)$$

$$\ddot{v} = Y_1 v|v| + Y_2 \ddot{\vartheta} + Y_3 r^2 \quad (6)$$

In Eqs. (4), (5) and (6), N_i , X_i , Y_i represent constant parameters which are estimated starting from the CETENA data.

These equations give a simplified expression of the more complex relations between the relevant quantities. Indeed, they could not be used in an "a

priori” model, but the simplification introduced by the use of these expressions is compensated by the constant parameters extracted with an interpolation algorithm. Thanks to these parameters, the model is able to learn from known data how the ship would behave answering to certain input.

Parameters extraction

The constant parameters needed to integrate the equation of motion (Eqs. (4), (5) and (6)) were obtained from sea trials results collected in CETENA files. To do that, it is useful to consider as an example Eq. (4), which can be rewritten to highlight the time-varying terms.

Given that in the CETENA data all the time-varying quantities are known, while the constant parameters are unknown, it is possible to rewrite Eq. (4) in a matrix form:

$$\begin{bmatrix} \ddot{\theta}(t_1) \\ \dots \\ \ddot{\theta}(t) \\ \dots \\ \ddot{\theta}(t_e) \end{bmatrix} = \begin{bmatrix} \sin(\delta(t_1)) u(t_1)^2 & r(t_1)^2 & (p_1(t_1)^2 - p_2(t_1)^2) \\ \dots & \dots & \dots \\ \sin(\delta(t)) u(t)^2 & r(t)^2 & (p_1(t)^2 - p_2(t)^2) \\ \dots & \dots & \dots \\ \sin(\delta(t_e)) u(t_e)^2 & r(t_e)^2 & (p_1(t_e)^2 - p_2(t_e)^2) \end{bmatrix} * \begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix} \quad (7)$$

In this expression, the number of equations is much greater than the number of unknown terms, so that the system can be solved in terms of a least square interpolation performing a pseudo-inverse of the matrix (*pinv* function). The same procedure can be repeated for the other two equations of motion (Eqs. (5) and (6)) in order to calculate all the constant parameters needed (i.e. N_i , X_i , Y_i). The maneuver described in CETENA records used for this calculation was the turning circle (Figure 2).

Initial conditions

Once the parameters have been estimated, it was possible to integrate Eqs. (4), (5) and (6) to obtain a simulation of the ship behaviour. The integration method used was the Forward Euler method. The simulator needs as an input the pilot’s instruction (i.e. rudder angles and propellers speed) and the initial conditions (i.e. starting coordinates and heading and starting linear and rotational speed). The speed information could be obtained by deriving coordinates and heading with respect to the time. Anyway, those data are affected by the error caused by the GPS and the compass sensitivity, so that a correct estimation of the initial speed is not trivial. A simple incremental ratio would give an unstable result, considering two consequent points, because of the measures noise. A more robust way to proceed is using a polynomial

interpolation to approximate the position and the heading data, in order to filter out such spurious effects.

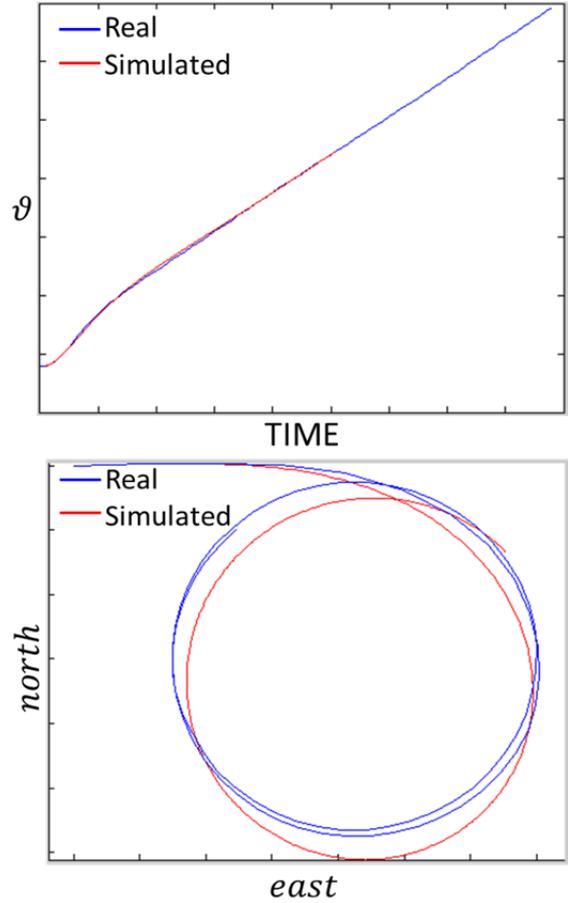


Figure 2: Turning circle: simulation time 500 s.

Validation

The best way to validate the aforementioned algorithm is to perform simulations of known data, in order to compare the simulated results with real ones. This was done with CETENA data contained in the Manoeuvring Booklet. The IMO standard manoeuvres were simulated by giving to the simulator the same values of rudder angles and propeller rpm (round per minute) used in real manoeuvres. The real data (position and heading) recorded by the sensors used during the sea trials were then compared with the simulator output. In Figure 2, Figure 3 and Figure 4 the results of this comparison are shown for turning circle, zig-zag and Williamson turn manoeuvres respectively. The results are surprisingly good if we take into account the simplicity of the mathematical/physical model used and the few information required to build it (i.e. the set of data collected during the sea trials).

To better prove the reliability of the simulator, a further validation was performed by using the data recorded in the VDR of Costa Concordia during the week before the shipwreck. It must be noted that the real time simulator had been realized for short-term simulation, and that the final goal was the simulation of the manoeuvres ordered by the Master of the Costa Concordia just before the impact with the rocks.

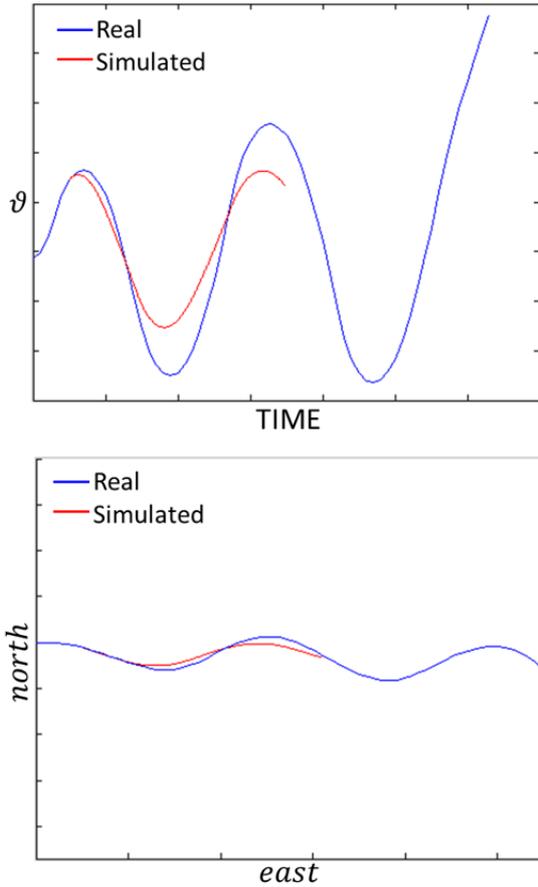


Figure 3: Zig-Zag: simulation time 300 s.

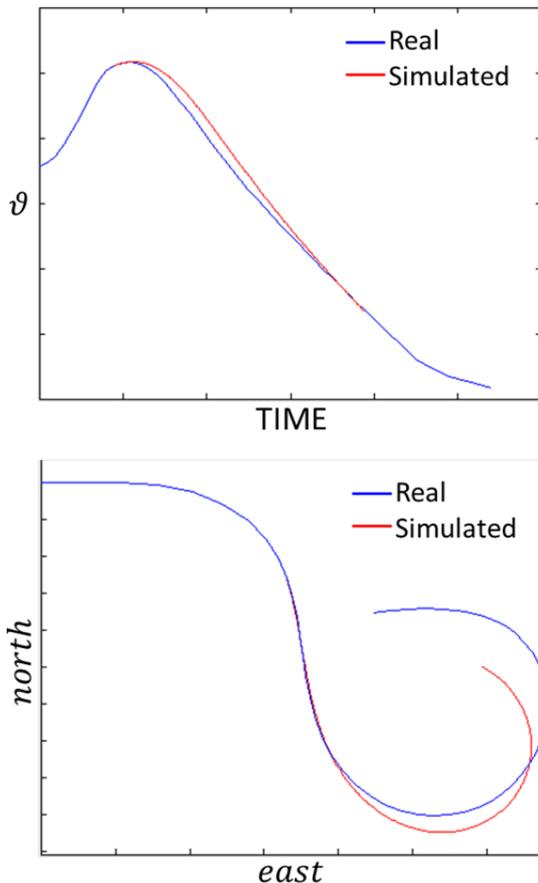


Figure 4: Williamson turn: simulation time 300 s.

Three events of the navigation recorded in the VDR were selected for validation. We will call them Palamos Turn, Palma Zig-Zag, Giglio Zig-Zag, because of the site in which they occurred and the manoeuvre typology. It is worth to underline that Giglio Zig-Zag manoeuvre ends with the impact against the rocks. Comparison results are reported in Table 1, in terms of position and heading errors. The errors were calculated at different simulation times.

Palamos Turn			
Sim. time [s]	east error [m]	north error [m]	Heading error [°]
30	1.6	0.3	-0.76
60	15.9	2.9	-3.42
120	82.2	-40.8	-14.60
Palma Zig-Zag			
Sim. time [s]	east error [m]	north error [m]	Heading error [°]
30	-0.5	-0.8	-0.07
60	-1.5	-2.1	-0.09
120	-4.1	-7.6	-0.21
Giglio Zig-Zag			
Sim. time [s]	east error [m]	north error [m]	Heading error [°]
30	5.3	5.2	-0.26

Table 1: Simulation errors.

As can be seen, the simulation results are reliable for a simulation time of about 60 s, while the error increases as the simulation time reaches 120s. This can be justified by the fact that in a short time simulation, environmental conditions, neglected in the model, play a second order role in ship behaviour, while become much more important as the simulation time increases. Furthermore, the integration method used is really simple in order to reduce simulation time, then the effects of the errors cumulate as simulation time increases.

4. RESULTS

In the case of Giglio Zig-Zag the simulated position of the ship at 21.45.11 (the impact time and the end of the simulation time) is considered. This simulation begins 30 seconds before the impact, that is, more or less the instant at which the Master saw the rocks and made the extreme attempt to avoid the impact by ordering a “Zig-Zag” manoeuvre. Apparently, the helmsman did not understand the orders of the master and made a mistake, by putting the rudder hard to starboard instead of hard to port as he was ordered. At this point, we repeated the simulation by giving to the simulator the correct position of the rudders as ordered. The timing of the order and the order itself were found in the recordings of the bridge audio, which was stored in the VDR, according with IMO Regulation. The result is shown in Figure 5, where the blue line represents the real position of the ship at the impact time and the

green one indicates the simulated position of the ship without the error of the helmsman. The black square in the figure represents the rock. The quite surprising conclusion is that the impact point at the left side of the ship in the simulation is more or less 10 m far from the rock. Taking into consideration the sea bottom shape at the accident site, this distance would be enough to allow the ship to pass next to the rock without any damage.

To complete the discussion, we also considered the error worst case. The maximum observed value of the error after 30 s of simulation is about 7 m: in this case, we would still have an impact but the actual impact point would have been about 18 m behind the actual one. It is our opinion that, in this worst case, at least one of two engine rooms (more likely both of them) would have been left unscathed by the accident, and the ship would have been capable to stay afloat. In this case, the final consequences on the ship and the passengers would have been really different.

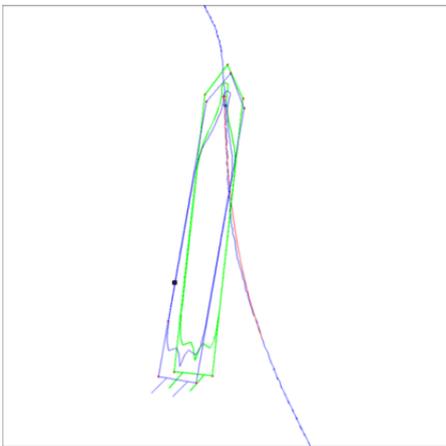


Figure 5: Real (blue) and simulated (green) position of the ship at the instant of the impact.

5. CONCLUSION

A simple time domain short-term simulator of the motion of a ship in free water has been presented. The simulator is based on a quite simple “black-box” model that does not require the knowledge of structural and mechanical characteristics of the ship. The position of the ship and its orientation versus time are calculated by using a mathematical model in which ten free parameters are used. These parameters are bound to the manoeuvrability characteristics of the ship, and can be obtained by using the recorded results of sea trials performed with the ship before its delivery.

Despite the great number of papers in literature dealing with the problem of simulating the manoeuvres of a specific ship, just a few of them deal with the problem of a time domain algorithm able to simulate, at least for a short time, the effect of propellers and rudders during a manoeuvre.

Our simulator could be used for pilots and masters training, in autopilot systems, in emergency manoeuvre foreseeing and so on. Further improvements will be introduced in the model in order to take into account the effects of wind and stream.

In the second part of the paper, the procedure has been used to simulate the Costa Concordia cruise ship behaviour, in order to study the effect of the allegedly wrong manoeuvre performed by the helmsman just before the impact. The result was that if the helmsman would properly execute the Master’s orders, the impact could have been avoided or, at least, by considering the worst case, its effects could have been much less devastating. Finally, to definitely confirm these conclusions, the Concordia’s twin ship, Costa Serena, could be used to perform a sea trial in which the manoeuvres with and without the helmsman’s error should be executed. The different tracks recorded during the test could then be compared to evaluate the actual consequences of the helmsman’s error on the Costa Concordia shipwreck.

6. ACKNOWLEDGEMENT

The authors are grateful to CODACONS, an Italian consumers’ rights association that is supporting some Costa Concordia survivors in the trial, for financial support and legal assistance.

REFERENCES

- Damitha Sandaruwan, Nihal Kodikara, Chamath Keppitiyagama and Remy Rosa, 2010, “A Six Degrees of Freedom Ship Simulation System for Maritime Education” , *The International Journal on Advances in ICT for Emerging Regions* , pp 34-47
- Daldoss L.; P. Gubian; Quarantelli M. (2001). Multiparameter Time-Domain Sensitivity Computation. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I. FUNDAMENTAL THEORY AND APPLICATIONS* p. 1296 - 1307 Vol. 48
- Gatis B. and Peter F.H., 2007, "FAST-TIME SHIP SIMULATOR.". *Safety at Sea*.
- Gear C.W. - *Numerical Initial Value Problems for Ordinary Differential Equations*, Prentice-Hall, 1971
- IMO - International Maritime Organization. Resolution A.861(20), 1997, “Performance standards for shipborne Voyage Data Recorders (VDRs)”. Adopted on November 27.
- Ishiguro T., S. Tanaka, and Y. Yoshimura, 1996, “A Study on the Accuracy of the Recent Prediction Technique of Ship’s Manoeuvrability at an Early Design Stage,” In *Proceedings of Marine Simulation and Ship Manoeuvrability*, M.S. Chislett, ed., Copenhagen, Denmark
- ITTC, 1996, International Towing Tank Conference, 1996, "Manoeuvring Committee - Final Report and Recommendations to the 21st ITTC", *Proceedings of 21st ITTC, Bergen - Trondheim, Norway, Vol. 1*, pp. 347-398.
- ITTC, 1999, Recommended Procedures-Testing and Extrapolation Methods -Manoeuvrability Validation of Manoeuvring Simulation Models, available at [http://www.simman2008.dk/PDF/7.5-02-06-03%20\(simulation%20models\).pdf](http://www.simman2008.dk/PDF/7.5-02-06-03%20(simulation%20models).pdf)
- ITTC, 2002a, International Towing Tank Conference, 1999, "Manoeuvring Committee – Final Report and Recommendations to the 22nd ITTC", *Proceedings of 22nd ITTC, Seoul/Shanghai. International Towing Tank Conference, 2002*.
- ITTC, 2002b, "Esso Osaka Specialist Committee – Final Report and Recommendations to the 23rd ITTC", *Proceedings of 23rd ITTC, Venice, Italy*.

- Maffezzoni P., P. Gubian "A New Unsupervised Learning Algorithm for the Placement of Centers in a Radial Basis Function Neural Network (RBFNN)", Proceedings of 1995 International Conference on Neural Networks, Perth, Western Australia, Nov. 1995, Vol. 3, pp. 1258—1262
- Maimun A., Priyanto A., Muhammad A.H., Scully C.C., Awal Z.I., 2011, "Manoeuvring prediction of pusher barge in deep and shallow water", *Ocean Engineering* 38, 1291–1299
- Maimun A., Priyanto A., Rahimuddin A.Y., Sian Z.I., Awal, Celement C.S., Nurcholis, Waqiyuddin M., 2013, "A mathematical model on manoeuvrability of a LNG tanker in vicinity of bank in restricted water", *Safety Science* 53, 34–44
- Mohd N.C.W., Wan N.W.B. and Mamat M., 2012, "Ship Manoeuvring Assessment By Using Numerical Simulation Approach" *Journal of Applied Sciences Research*, 8(3): 1787-1796.
- Oltmann, P., 1996, "On the Influence of Speed on the manoeuvring behaviour of a container carrier", In Proceedings of Marine Simulation and Ship Manoeuvrability, Copenhagen, Denmark, ISBN 90 54108312
- Piccinelli M. and Gubian P., 2013, "Modern Ships Voyage Data Recorders: a Forensics Perspective on the Costa Concordia Shipwreck" *Digital Investigation* 10, Elsevier.
- Poggio T., F. Girosi "Networks for Approximation and Learning", Proceedings of the IEEE, vol. 78, no. 9, September 1990, pp. 1481-1497
- Revestido E. et al, 2011, "Parameter Estimation of Ship Linear Maneuvering Model", Proceedings of Ocean 2011, pp. 1-8, 6-9, Santander, Spain.
- Sutulo S., Moreira L., Guedes Soares C., 2002, "Mathematical Models for Ship Path Prediction in manoeuvring simulation systems", *Ocean Engineering*, Vol. 29, Pergamon Press, pp.1-19.
- Shyh-Kuang U., David Lin, Chieh-Hong Liu., 2008, "A ship motion simulation system." s.l.: Springer-Verlang, Virtual Reality. pp. 65–76.

AUTHOR BIOGRAPHIES



PAOLO NERI was born in Palermo, Italy, and he earned his Master Degree in Mechanical Engineering at the University of Pisa in 2012. He is now attending the second year of the PhD course at University of Pisa. His main

research interests include Physical systems simulation, Experimental Modal Analysis, Machine Dynamics and especially the development of robot-assisted automatic procedures for mechanical systems dynamic characterization.

His e-mail address is: paolo.neri@for.unipi.it.



MARIO PICCINELLI was born in Lovere, Italy, and earned his Master's Degree in Electronic Engineering at the University of Brescia, Italy, in 2010. He is now a Ph.D. candidate in Information Engineering with a thesis about Digital

Forensics. His main research interests include the extraction and analysis of data from modern embedded devices, such as smartphones or eBook readers. He is also working as a consultant for both

prosecutors and lawyers for matters related to the analysis of digital evidence, and was appointed as private consultant in the Costa Concordia shipwreck trial. His e-mail address is: mario.piccinelli@gmail.com and his Web page can be found at <http://www.mariopiccinelli.it>



PAOLO GUBIAN received the Dr. Ing. degree "summa cum laude" from Politecnico di Milano, Italy, in 1980. He consulted for ST Microelectronics in the areas of electronic circuit simulation and CAD system architectures. During this period he worked at the design and implementation of ST-SPICE, the company proprietary circuit simulator. From 1984 to 1986 he was a visiting professor at the University of Bari, Italy, teaching a course on circuit simulation. He also was a visiting scientist at the University of California at Berkeley in 1984. In 1987 he joined the Department of Electronics at the University of Brescia, Italy where he is now an Associate Professor in Electrical Engineering. His research interests are in statistical design and optimization and reliability and robustness in electronic system architectures. His e-mail address is: paolo.gubian@unibs.it and his Web-page can be found at <http://www.ing.unibs.it/~gubian>.



BRUNO NERI was born in 1956 and received his "Laurea" degree "cum laude" from University of Pisa in 1980. In 1983 he joined the Department of Information Engineering of the same University, where he is full Professor of

Electronic since 2000. In recent years his research activity has addressed the design of radio-frequency integrated circuits for mobile communications and for biomedical applications. Presently he is Technical Consultant in the lawsuit for the shipwreck of Costa Concordia cruise ship. Bruno Neri is co-author of more than 100 papers published in peer-reviewed Journals and Proceedings of International Conferences.

His e-mail address is: b.neri@iet.unipi.it.

INTEROPERABILITY IN CO-SIMULATIONS OF MARITIME SYSTEMS

Christoph Dibbern
Axel Hahn
Carl von Ossietzky Universität Oldenburg
Ammerländer Heerstr. 114-118
D-26129, Oldenburg, Germany
E-Mail: {dibbern, hahn}@wi-ol.de

Sören Schweigert
OFFIS - Institut for Information Technology
Escherweg 2
D-26131, Oldenburg, Germany
E-Mail: soeren.schweigert@offis.de

KEYWORDS

Distributed Simulation; Maritime Traffic Simulation; Interoperability; Model Transformation; Semantic Model; Conceptual Architecture; Layers for Simulation-based Analysis; HLA; RTI;

ABSTRACT

Powerful, flexible and cost-effective analyses of maritime systems can be conducted in an appropriate way with distributed simulations. Co-simulation components require interoperability like specified by the IEC TC/65/290/DC to insure information exchange and synchronization. Therefore this paper describes a way to insure the interoperability of co-simulations of maritime systems. User- and technical requirements for these co-simulations are derived to create a deeper understanding for especially the necessity of a common semantic model. Furthermore, six integration layers are presented for co-simulation-based analysis. They are derived from IEC TC/65/290/DC and the conceptual architecture approach for co-simulation systems. The paper introduces a semantic model approach to reflect the requirements formal description of the simulation model, observability, controllability and interoperability. The model is used to configure the inter-process communication of the used co-simulation architecture. This is done by an automatic model transformation from the semantic model to the platform specific implementation.

INTRODUCTION

Seafaring has always been a joint undertaking between humans and their technology. The reliability of the technical equipment and its correct usage ensure safe travelling. This is still true with the implementation of eNavigation technology.

The eNavigation implementation process is accompanied by IMO's NAV and COMSAR sub-committees, as well as the International Hydrographic Organization (IHO) and the International Association of Lighthouse Authorities (IALA). These institutions did a comprehensive gap analyses as part of their development of a joint implementation plan for eNavigation. In a ten year survey (Gale 2007) investigated the causes of collisions and groundings, in which human error was the primary cause with 60%.

Therefore the gap analysis of the IMO addresses numerous aspects of human machine interaction (IMO 2012), e.g. absence of structured communication to report incorrect operation of both shipboard and/or shore-based systems together with a lack of intuitive human-machine interface for communication and navigation means. Furthermore, the analysis revealed that existing performance standards are not applied or are missing such as guidelines for usability evaluation (IMO 2012). This requires that equipment providers have to do a comprehensive usability and risk assessment of their products. IMO MSC Circular 878 states: "A single person's error must not lead to an accident. The situation must be such that errors can be corrected or their effect minimized. Corrections can be carried out by equipment, individuals or others."

Cognitive Simulation Based Test Bed

To fulfill the announced IMO requirements the authors propose a cognitive simulation based test bed for human machine interaction, to provide a test bed for experiments during early design phases of an eNavigation System like an Integrated Navigation System (INS).

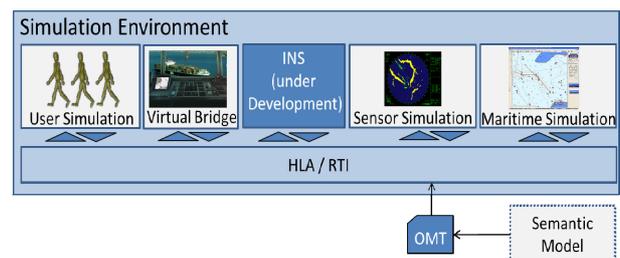


Figure 1: Simulation-based analysis of human machine interaction

Figure 1 shows the proposed test bed containing the system under development that is tested in a co-simulation environment: A maritime traffic simulation provides the context for the experiments. A sensor simulation uses the traffic model to generate AIS and radar signals. The data is streamed to the *INS* for testing purposes. We use a simulated user (a virtual crew member implemented by a cognitive simulation) (Wortelen et al. 2006) which interacts with the not yet ready *INS* via another system: the virtual bridge simulator. It provides the 3D Environment of the ship

bridge in which the new *INS* is used and implements the human machine interaction (Puch et al. 2012).

Each simulation component is implemented as a separate simulation that must be synchronized in terms of time and communicate the dynamic information like vessel position, headings etc. with all the others.

All simulators must have the same understanding of their shared environment in order to carry out a joint simulation. Therefore a vessel within the *Maritime Traffic Simulation* is the same entity of a vessel as in the *Sensor Simulation*, *Integrated Navigation System* and the *User Simulation*. The common *Semantic Model* shown in Figure 1 is used to describe all items and the shared environment of all simulators.

In our case, this is technically archived by using a High Level Architecture (HLA) compliant implementation (Noulard et al. 2009) like described in Läsche et al. (2013). This implementation requires a formal description of the communicated data, in form of an Object Model Template (OMT).

INTEROPERABILITY REQUIREMENTS FOR THE CO-SIMULATION

To define the co-simulation requirements first bottom-up and top down practical requirements are identified. Secondly the authors take the IEC requirements on interoperability into account and finally the authors derive requirements from layered integration architecture for co-simulation systems.

Practical Requirements

Bottom-up it is necessary to find the technical requirements on the system to derive a suitable *Semantic Model* and architecture so that the user requirements can be fulfilled. In addition, the analysis of maritime systems generates user requirements.

These requirements of co-simulations of maritime systems can be emphasized with a top-down, problem-oriented analysis. Its goal is to derive the user requirements which have to be satisfied so that the co-simulation is able to support the usage scenarios. This focusses on the question: Which user requirements have to be satisfied to support the needed scenarios for analyzing purposes? The requirements of the usage scenario “analysis for maritime systems” can be split into two groups (s. Figure 2). The related user requirements are visualized as inner rectangles and are derived during the requirements analysis. The first group comprises usage scenarios like the Simulation Description and the automatic *Formal Analysis* of offshore operations as described in Läsche et al. (2014). The *Simulation Description* is e.g. necessary to describe the maritime environment like buoys and havens on waters like the Weser. Moreover static vessel characteristics like the power of its engine and its maximum draught can be described. An example for the automatic *Formal Analysis* is the analysis of hazard events like an injured person at sea as described in Läsche et al. (2014). Their approach allows assessing

the reasons for hazards. Relevant reasons can be e.g. environment factors (wind, waves etc.), broken equipment and human failures. Therefore a complete and correct *Formal Description* (e.g. of the agents behavior and the environment) is required. Because of that the *Semantic Model* has to provide all data types and parameters for the *Formal Description*. The second group includes the requirements regarding the *Controllability*, *Observability* and *Interoperability*. *Controllability* means the possibility to influence the behavior of the simulation components as well as the whole co-simulation with commands like start and stop. Moreover, the controllability also covers the injection of manual and automatic failures like described in Läsche et al. (2013). Furthermore, the co-simulation has to be *observable* because the user has to be able to monitor the internal states of the simulation components. An example is the logging of GPS positions of the simulated vessels for collision detection.

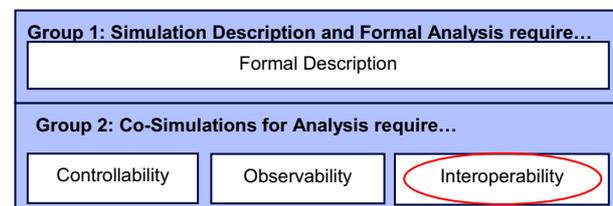


Figure 2: Problem-oriented View for the User Requirements

The user requirement *Interoperability* and its technical requirements are explained in the following chapters.

Interoperability Requirements

According to IEC (2002) a system is interoperable if: “The application data, their semantic and application related functionality of each device is so defined that, should any device be replaced with a similar one of different manufacture, all distributed applications involving the replaced device will continue to operate as before the replacement, but with possible different dynamic responses.” The insurance of the *Interoperability* is necessary for the simulation consistency during runtime. Table 1 shows six *System features* which are necessary for an interoperable system according to IEC (2002). This means, *Interoperability* stands for the interaction of two or more systems which fulfill all of the listed system features.

Summarized, technical requirements for the *Interoperability* are according to the IEC a common *Communication Protocol*, *Communication Interface Data Access* and *Data types*. These are necessary to provide the network infrastructure of distributed systems. Furthermore, the definition of *Parameter Semantics* and dependency and consistency rules is required to insure the *Interoperability* of the system (s. *Application Functionality*). The IEC *System features* help to derive the technical requirements for the co-simulation integration.

Table 1: System Features for *Interoperability* (IEC 2002)

System feature	Description
Application Functionality	“This feature is defined by specifying the dependencies and consistency rules within the Functional Elements. This is done in the data description part or in a separate behavior section.”
Parameter Semantics	“This feature is defined by the characteristic features (parameter attributes) of the application data this can be data name, data descriptions, the data range, Substitute value of the data, default value, persistence of the data after power loss and deployment.”
Data Types	“The data type of the object attributes or block data input, data output or parameter defines this feature.”
Data Access	“This feature is defined by the object operation definition or the access parameter attributes of the block data input, data output and parameters.”
Communication Interface	“This feature is defined by the communication service definition of application layer including the services and the service parameters. Additional mapping mechanisms can be necessary. The dynamic performance of the communication system is part of this feature.”
Communication Protocol	“This feature is defined by all protocols of layer 1 to 7 of the OSI (Zimmermann 1980) reference model, i.e. from the physical medium access to the application layer protocol.”

Table 2: Layers for Co-Simulations of Maritime Systems based on (Schütte 2013)

Category of TRs	Layer	Artifact
Control of the Simulation	6) CONTROL	Controlling Applications
Modelling / Scenario Composition	5) COMPOSITION	Scripts to configure the system to simulate e.g. a ship’s bridge
Decomposition	4) SCENARIO	Distributed Models in Domain Specific Simulators
Definition of the Semantics	3) SEMANTIC	Model for the Interoperability , Observability and Controllability
Configuration of the RTI	2) SYNTACTIC	RTI-Configuration File
Connection by Infrastructure	1) TECHNICAL	RTI (IPC, Time-Synchronization), Observer

Co-Simulation Integration

The conceptual architecture for simulation-based analysis from Schütte (2013) includes six layers (s. Table 2). The following description focusses on the first three layers because they are necessary to insure the *Interoperability* (Schütte 2013). The authors

extended the conceptual architecture by technical requirement (TR) categories and on the right side of the table, the artifacts per layer for co-simulations of maritime systems. The **first** layer provides the network connection and infrastructure. Therefore the system features *Communication Protocol*, *Communication Interface* and *Data Access* from Table 1 can be assigned to this layer. The associated artifact is a HLA-based Runtime Infrastructure (RTI) to provide an inter-process communication (IPC) as well as a time-synchronization. These are sensible to synchronize the distributed simulation components regarding their simulation states so that their communication is insured (Läsche et al. 2013). Another element on the first layer is the *Observer*. It is sensible for the infrastructure because it helps to analyze maritime systems through the recognition of predefined events during co-simulations (Läsche et al. 2013). Therefore, *Observer* are only able to listen on the HLA communication interface in this case (Läsche et al. 2013). The artifact on the **second** layer is necessary for the configuration of the RTI. This is a configuration file which contains the exchangeable data types, their units and other necessary information. This means, the system feature *Data Types* can be assigned to this layer. The **third** layer has to contain all data types and parameters with their semantic to insure the correctness and completeness of the co-simulation states. This helps to insure the communication respectively the data exchange between the co-simulation components. Furthermore, the layer includes all data types and parameters which are necessary to describe the possible behavior of the simulation components (commands like start, stop, pause etc.) so that it fulfills the system feature *Application functionality*. Therefore it includes the *Parameter Semantics* and the *Application Functionality* from Table 1. It follows that the satisfaction of the requirements from the first three layers insures the *Interoperability* of a co-simulation as defined by the IEC.

The artifact is the *Semantic Model* for both system features with all necessary data types and the *Parameter Semantics*. Summarized, it can be defined as follows:

The *Semantic Model* insures the *Interoperability*, *Observability* and *Controllability* of the simulators. Furthermore, it contains entities and parameters to describe static information like environments.

The design rationales for the creation of the *Semantic Model* are described in the following chapter.

The **fourth** layer includes the domain specific scenario simulations. A scenario is runnable by an individual simulator to simulate e.g. that a human is able to move on a ships bridge. The **fifth** layer is responsible for the scenario configuration. As described, each simulator has its own scenario. The scenario in a co-simulation is a composition of each scenario of each simulator to simulate e.g. ship’s bridge in combination with the maritime traffic on waters like the Weser.

The **sixth** layer includes applications to control the whole co-simulation and its components. They use the defined data types and parameters of the third layer which are necessary to fulfill the system feature *Application functionality*. The *Controllability* is insured if the requirements of all six layers are fulfilled.

Conclusion for the Co-Simulation Integration

Overall the following technical requirements on the simulation components are primary relevant for the *Interoperability* according to Schütte (2013): The simulation components have to be connectable to a *RTI* (s. layer 1 and 2 of Table 2). They must have a common understanding of the time so that the co-simulation states can be resumed to a defined point in time (s. layer 1). The simulation components use a common *Semantic Model* for their communication (s. layer 3). Their *Control Unit* uses unique, interoperable control commands like *start*, *stop* (s. layer 3). The *Control Unit* is able to synchronize the step size of each simulator as defined. This is necessary to insure the consistency of the co-simulation states. In addition, data type and range matching has to be performed (Schütte 2013). The data types have to be matched so that it is insured that the internal data types of the simulation components are compatible with the data types which are specified in the common *Model*. Range matching is the check of the data type ranges. This is necessary to insure that the possible value ranges have the required level of granularity.

DESIGN RATIONALES FOR THE MODEL

This chapter focusses on the design rationales for the model from the third layer (s. Table 2). The model design is based on API design rationales as described in Bloch (2006) and Tulach (2008) to increase its reusability and quality. The following design rationales are explained below: the use of standards, object oriented design and transitive dependencies.

The use of established **standards** (ISO, IEC etc.) helps to cover the correctness and completeness of the model. This allows to consider e.g. the correct description of the environment with the ISO19125 (ISO 2004) which is a specification for geographic information meta data. Moreover the appropriate use of standards increases the probability that the model can be reused for projects with the focus on maritime risk and efficiency analysis. Furthermore the model is based on standards for the system engineering like the Department of Defense Architectural Framework (DoDAF 2011). This specifies e.g. the vocabulary for the description of the DoDAF meta models. All used meta-model standards are explained more detailed in the next chapter.

The **object oriented design** with its hierarchical structure allows an effective reuse of model elements. This design comprises e.g. the use of polymorphism and encapsulation. The practical use for polymorphism is e.g. to be able to describe that there are several types to specify a point in space that could be either a two

dimensional space like it is used in a traffic simulation on the water plane, or a three dimensional space required in an underwater simulation.

Encapsulation means in this case that the related data types and parameters are logical organized in model parts (s. Figure 3). Therefore, the encapsulation helps to consider the separation of concerns principle (Gurp and Bosch 2003). For example, the *Traffic System* has no information about elements defined in the *Sense* model. As consequence, changes in the *Sense* model have no impact on the *TrafficSystem* and vice versa. In addition this covers the need of a slim model. The **transitive dependencies** of the model parts ensure that all types used in a more general package are available, even if the dependency is not modeled explicit.

THE SEMANTIC MODEL

This chapter focusses on the structure and used standards of the model from the third layer (s. Table 2). The basic model structure is visualized in Figure 3. It includes two layers. The top layer contains model parts with general data types and concepts. The layer below is for the two model parts with domain specific data types. The meta-model standards *ISO19125*, the *SensorML* and *Safety ISO26262* are used for this layer. The standards are explained together with the layers below.

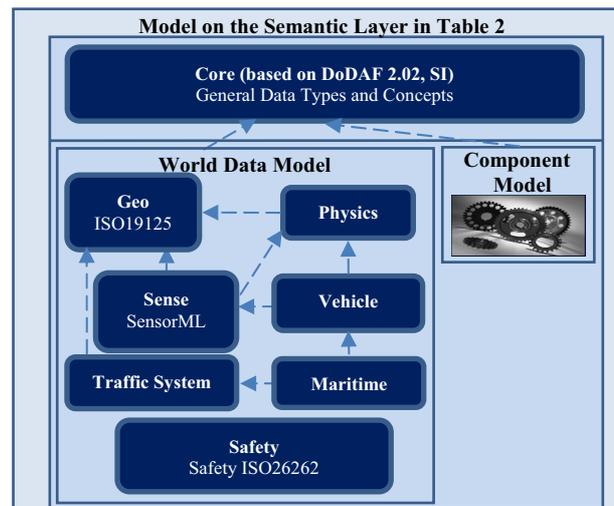


Figure 3: Structure of the Model with Transitive Dependencies

The *Core* is an elementary, object oriented model part with general data types and concepts. *Concepts* are empty shells. They cannot be instantiated and have to be specified in sub models. In addition, they are useful to categorize data types and parameters. The concept *Vehicle* shows e.g. the relation between *Vessels*, *Cars* and *Amphibious Crafts*. The *Core* is oriented on DoDAF 2.02 (DoDAF 2011) and ‘The International System of Units’ (National Institute of Standards and Technology 2008). Therefore the *Core* contains general data types like *SIUnit*. DoDAF comprises general concepts like ‘*Activity*’, ‘*Capability*’ and ‘*Performer*’. A ‘*Performer*’ comprises an ‘*Activity*’ as well as a

‘Capability’ (DoDAF 2011). An example for a ‘Performer’ is e.g. the captain of a vessel which can be simulated as a user like explained in Figure 1.

The *World Data Model* (WDM) and the *Component Model* (CM) are in the layer below. The CM contains all entities and parameters which are necessary for the control of the co-simulation components, like explained for the third layer of Table 2.

Data Types to Describe the Current Situation

The WDM is the *Semantic Model* for the description of data types and parameters which are necessary for the communication of the current situation between the simulation components. It is used to describe all necessary data types and parameters to integrate simulators which are necessary for the representation of the Simulation Environment like a shipping area in the case of the example mentioned in the introduction.

The *Geo* part follows the Simple Feature Access which is specified by the ISO19125. It is used to describe geographical data. Therefore it comprises data types like ‘GeoPoint’, ‘Curve’, ‘Surface’ and ‘Polygon’ that could be both two- or three-dimensional. Furthermore, the model needs to contain data types and parameters for the integration of sensor simulations and physical simulations which are included in the *Sense*, respectively the *Physics* part. In the case of maritime systems it is e.g. required that the *Sense* part includes data types and parameters for the integration of sensor simulations like GPS so that the Maritime Traffic Simulator is able to communicate the vessel positions to a radar simulator. Because of that the *Sense* part is oriented on the *SensorML* (Open Geospatial Consortium Inc. 2007). The Sensor Modelling Language is a XML based standard which is used to describe sensors as well as their measurement process. This standard focusses on the description of sensing devices. It also includes models to describe observations of these devices. The *Vehicle* part is necessary to integrate Vehicle simulators like for Vessels with data types like draught and length of a vessel. The *TrafficSystem* part includes data types like ‘WayPoint’ and ‘Trajectory’. The *Maritime* part is for the integration of Maritime Simulations with data types like ‘Wave’ and ‘Ocean’.

Safety relevant Data Types in the Maritime Context

In addition, the *Safety* part contains data types and parameters for the recognition of hazards and failures so that rare events can be recognized like described in Läsche et al. 2013. Therefore the *Safety* part is oriented on the *Safety ISO26262* (ISO 2012). This is a specification for functional safety in the automotive area which includes, among others, an optimized vocabulary for its scope (Hagel 2011; Esposito 2010; LDRA 2011). The *Safety ISO26262* is based on the IEC 61508 and contains many general safety relevant elements for the automotive domain. The IEC 61508 is a specification for safety related systems (LDRA 2011;

exida 2006). The *Safety ISO26262* can be used for scenarios regarding safety analysis for maritime systems. Because of that it is sensible to derive the relevant data types in the context of safety analysis for maritime systems. For example, data types like hazards and failures can be used to simulate that a failure consists of a concatenation of events. In the case of automotive, a hazard is defined as (Esposito 2010): “A potential source of harm. Harm is a physical injury or damage to the health of people“. A failure can be defined as (Esposito 2010): “The termination of the ability of an element or an item to perform a function as required.” In the case of maritime systems, an example for a failure is the grounding of a vessel in a curve because of an occurred hazard like the harm of the vessel engine).

The *Semantic Model* is described using the Eclipse Modelling Framework (EMF) (Steinberg et al. 2008) to meet the user requirement for a *Formal Description*. Therefore the EMF ability to generate a valid implementation out of the *Formal Description* can be used. Furthermore, this allows the use of the EMF tool chain for further development.

SEMANTIC MODEL INTO OBJECT MODEL TEMPLATE TRANSFORMATION

This chapter focusses on the transformation of the *Semantic Model* from the third layer of Table 2 into an OMT compliant format. The *Model* describes the common data types and concepts of the various simulators involved in a joint simulation. It is constructed object-orientated, on the basis of the requirements of the previous chapter. A strong use of polymorphism ensures among others the reusability, as described in the chapter above. In contrast, most RTIs are based on a much simpler data model, e.g. without multiple inheritance. This also applies to the HLA implementation used by Läsche et al. (2013) which only allows the use of fundamental types (integer, float, etc.) as well as their array representation. Moreover, the HLA implementation utilizes XML files to describe the OMT. Therefore model driven architecture (s. Brambilla et al. 2012; Stahl et al. 2007) techniques are used to generate the platform specific OMT files from the platform-independent *Semantic Model*. This also eliminates the need to maintain two versions of the same model which is the EMF file of the *Semantic Model* and XML file in the case of the OMT. Figure 4 shows a possible modelling of a datave called *Pose*.

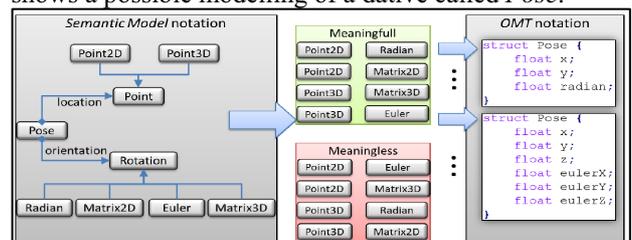


Figure 4: Breakdown of complex data structures to a set of fundamental data types

The *Pose* contains a location (*Point*) as well as an orientation (*Rotation*). Both types can be specialized for a two or three dimensional simulation, like mentioned in the introduction (upper side of the figure).

A *Pose* can exist in at least eight different variations, like displayed on the lower left part of Figure 4, where four meaningless combinations (mixing of dimensions) and four meaningful combinations can be found. The premise is in this case, that an HLA Object cannot contain a complex data type. Therefore the last step, each of these combinations has to be decomposed into a set of fundamental data types (like displayed on the rightmost side of the figure) to fit the requirements of the platform specific OMT. Moreover, it involves that each object of the *Semantic Model* which constitutes a *Pose*, e.g. a vessel can also exist in at least eight different “flat” variations of a *Pose* like mentioned in Figure 4. In a more complex model this would lead to a huge number of possible variations for a single object, also containing a huge amount of meaningless combinations of attributes like in Figure 4.

The simulators involved in a joint simulation need to agree on a combination of attributes in reality since in most cases each data type need a special treatment. For example, a *Pose* is the combination of a *Point3D* as *location* and an *Euler* as *orientation*.

Manual Selection of Concepts

In addition, not all information available within the model is needed to couple two or more simulations. For example, the maritime simulation mentioned in the introduction does not require knowledge about the number of crew members on the vessel, since it does not affect its dynamic behavior.

Taking these arguments into account, the authors propose a manual selection of concepts and their attributes, to perform the step from the semantic layer to the syntactic layer of Table 2. As additional benefit this will also lead to a lower usage of network bandwidth, to synchronize the simulations.

The selection of Concepts and attributes is then done based on the *Semantic Model*, in a form like:

From Concept *Vessel* take the attribute *pose* as type *Pose*, from this *pose* take the attribute *location* as a *Point2D* and finally take the attribute *X* as the communicated value.

The formal description of the *Semantic Model* then may be used to detect the fundamental data type of the requested attribute. This allows the automatic generation of RTI-Configuration files, combining the *Semantic Model* as well as the selection of fundamental attributes. In case of the HLA-OMT Files mentioned in Läsche et al. (2013) this is done using a Model to Text transformation, based on the XPand Project of the Eclipse Modelling Tools (Seidl 2013).

Currently, the transformation from the semantic model into the HLA-OMT format is supported. In the future, other inter-process communication technologies, such

as CORBA (OMG 2012) or ZeroC-ICE (Spruiell 2011) should be supported.

CONCLUSION AND NEXT STEPS

This article analyses the *Interoperability* in co-simulations of maritime systems. Therefore the requirements for these distributed simulations are derived by a problem-oriented as well as technical-oriented approach in the first chapter. The main part is the combination of the IEC compatibility level matrix with the conceptual architecture for simulation-based analysis. The paper introduces architecture with six layers for co-simulation *Interoperability* which defines requirements for co-simulations of maritime systems. A central role has the *Semantic Model* because it helps to insure the satisfaction of the derived user requirements: *Formal Description*, *Observability*, *Controllability* and especially the *Interoperability*. Therefore the semantic layer contains the *Semantic Model*. Its structure is based on the described design rationales and standards. The automatic model transformation is necessary to build a useful bridge between the *Technical Layer* and *Semantic Layer*. Its practical use is to insure the platform specific implementations consistency to the semantic model. Furthermore this automatic process accelerates the implementations creation.

One of the next steps is to check additional standards like the S-100 data model (IHO 2010) which supports the integration of hydrographic data and applications. In this case, it will be analyzed regarding useful elements for the *Semantic Model* to improve its *Interoperability* for the integration of simulators of maritime environments. Furthermore, a case study will be used to evaluate the *Interoperability* of the co-simulation in the case of the integration of a maritime traffic simulation (MTS) and radar simulations. The goal is to extend the *Semantic Model* so that the necessary generated ship data of the MTS can be processed by components for sensor simulation and AIS data generation. This allows testing and improving sensor systems.

REFERENCES

- Bloch, J., 2006. How to design a good API and why it matters. *Companion to the 21st ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications - OOPSLA '06*.
- Brambilla, M., Cabot, J. & Wimmer, M., 2012. *Model-Driven Software Engineering in Practice: Synthesis Lectures on Software Engineering*, Morgan & Claypool Publishers.
- DoDAF, 2011. The DoDAF Architecture Framework Version 2.02. Available at: http://dodcio.defense.gov/Portals/0/Documents/DODAF/DoDAF_v2-02_web.pdf [Accessed March 3, 2011].
- Droste, R. et al., 2012. Model-Based Risk Assessment Supporting Development of HSE Plans for Safe Offshore Operations. , pp.146–161.
- Esposito, C., 2010. *Hands on the ISO 26262 Standard*, Napoli: Università di napoli Federico II. Available at: <http://www.critical-step.eu/index.php/meeting->

- documents/doc_download/78-training-presentation [Accessed February 10, 2014].
- exida, 2006. *IEC 61508 Standard for Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*, Sellersville, Pennsylvania. Available at: http://www.win.tue.nl/~mvdbrand/courses/sse/1213/iec61508_overview.pdf [Accessed February 1, 2014].
- Gale, C.H., 2007. Improving navigational safety. , pp.4–8.
- Gurp, J. Van & Bosch, J., 2003. *Separation of Concerns : A Case Study*, Groningen. Available at: http://www.researchgate.net/publication/2563874_Separation_of_Concerns_A_Case_Study/file/79e4150c04dcf162ad.pdf [Accessed February 13, 2014].
- Hagel, J., 2011. *Herausforderung Funktionale Sicherheit*, Sindelfingen: MBtech. Available at: http://www.mbtech-group.com/fileadmin/media/pdf/electronics_solutions/Herausforderung-funktionale-Sicherheit_Web.pdf [Accessed January 30, 2014].
- IEC, 2002. *TC65: Industrial Process Measurement and Control*,
- IHO, 2010. *S-100 Universal Hydrographic Data Model*, Monaco: International Hydrographic Bureau.
- IMO, 2012. *Report to the Maritime Safety Committee, NAV 58/WP.6/Rev.1*,
- ISO, 2004. *ISO 19125-2:2004 Geographic information - Simple feature access*, Geneva. Available at: http://www.ird.fr/informatique-scientifique/methodo/standards/normes_iso_ogc/iso_geo/afnor/ISO_19125-2_DE_2004-ANGLAIS.pdf [Accessed December 28, 2013].
- ISO, 2012. *ISO 26262-10:2012 Road vehicles - Functional safety*, Geneva.
- Läsche, C. et al., 2014. MODEL-BASED RISK ASSESSMENT OF OFFSHORE OPERATIONS. In *Proceedings of the 33rd International Conference on Ocean, Offshore and Arctic Engineering*. San Francisco, CA, USA: ASME.
- Läsche, C., Gollücke, V. & Hahn, A., 2013. Using an HLA Simulation Environment for Safety Concept Verification of Offshore Operations. In *27th European Conference on Modeling and Simulation*. Alesung, Norway.
- LDRA, 2011. *ISO 26262 the Emerging Automotive Safety Standard*, LDRA.
- National Institute of Standards and Technology, 2008. The International System of Units [SI] B. N. Taylor & A. Thompson, eds. , 9(7), p.97.
- Noulard, E., Rousselot, J. & Siron, P., 2009. CERTI , an Open Source RTI , why and how. In *Spring Simulation Interoperability Workshop*. San Diego. Available at: <http://oatao.univ-toulouse.fr/2056/> [Accessed November 3, 2013].
- OMG, 2012. Common Object Request Broker Architecture (CORBA). Available at: <http://www.omg.org/spec/CORBA/> [Accessed February 5, 2014].
- Open Geospatial Consortium Inc., 2007. *OpenGIS® Sensor Model Language (SensorML) Implementation Specification*, Huntsville.
- Puch, S. et al., 2012. Rapid Virtual-Human-in-the-Loop Simulation with the High Level Architecture. In *Proceedings of Summer Computer Simulation Conference 2012 (SCSC 2012)*. pp. 44–50.
- Schütte, S., 2013. *Simulation Model Composition for the Large-Scale Analysis of Smart grid Control Mechanisms*. Oldenburg: Universität Oldenburg.
- Seidl, C., 2013. The Eclipse Modeling Framework (EMF): A Practical Introduction and Technology Overview. Available at: http://st-teach.inf.tu-dresden.de/wiki/images/The_Eclipse_Modeling_Framework_%2528EMF%2529_-_Christoph_Seidl_%2528commented%2529.pdf [Accessed October 10, 2013].
- Spruiell, M., 2011. ZeroC Documentation. Available at: <http://doc.zeroc.com/display/Doc/Home> [Accessed February 6, 2014].
- Stahl, T. et al., 2007. *Modellgetriebene Softwareentwicklung: Techniken, Engineering, Management*, Heidelberg: dpunkt.verlag.
- Steinberg, D. et al., 2008. *EMF Eclipse Modeling Framework* 2nd ed., Boston: Pearson Education, Inc.
- Tulach, J., 2008. *Practical API Design: Confessions of a Java Framework Architect*, New York: Apress.
- Wortelen, B., Lüdtke, A. & Baumann, M., 2006. Integrated Simulation of Attention Distribution and Driving Behavior.
- Zimmermann, H., 1980. OSI Reference Model-The ISO Model of Architecture for Open Systems Interconnection. *IEEE Transactions on Communication, Vol . COM-28, No.4 April*, pp.425–432.

AUTHOR BIOGRAPHIES



CHRISTOPH DIBBERN received his M.Sc. in Information Technology in 2012 at the University of Applied Science Kiel and started to work as a software developer for Business Process Management Systems at the company ESN innovo GmbH afterwards.

Since 2013, he has been working at the University of Oldenburg within the project CSE (Critical Systems Engineering). His E-Mail address is: dibbern@wi-ol.de and the homepage of his group is <http://be.wi-ol.de/>.



AXEL HAHN is full professor at the University of Oldenburg and leads the working group Business Engineering and board member of the division Transportation at the research institute

OFFIS. His research topics are safety and efficiency in marine transportation systems. His E-Mail address is: hahn@wi-ol.de and the homepage of his group is <http://be.wi-ol.de/>.



SÖREN SCHWEIGERT received his Dipl. Inform. in 2009 at the University of Oldenburg and started to work at OFFIS in the research group CMS (Cooperative mobile systems) afterwards. Currently he is working within the project COSINUS

with focus on Maritime Simulations. His E-Mail address is: soeren.schweigert@offis.de and the homepage of his group is <http://www.offis.de/en/start.html>.

Simulation and Visualization for Training and Education

RECYCLING A DISCARDED ROBOTIC ARM FOR AUTOMATION ENGINEERING EDUCATION

Filippo Sanfilippo*, Ottar L. Osen† and Saleh Alaliyat†

*Department of Maritime Technology and Operations, Aalesund University College,
Postboks 1517, 6025 Aalesund, Norway. Email: fisa@hials.no

†Department of Engineering and Natural Sciences, Aalesund University College
Postboks 1517, 6025 Aalesund, Norway.

KEYWORDS

PLC, automation engineering education, robotic arm.

ABSTRACT

Robotics and automation technology instruction is an important component of the industrial engineering education curriculum. Industrial engineering and automation departments must continuously develop and update their laboratory resources and pedagogical tools in order to provide their students with adequate and effective study plans. While acquiring state-of-the-art manufacturing equipment can be financially demanding, a great effort is made at Aalesund University College to provide the students with an improved hands-on automation integration experience without major capital investments. In particular, a strategy that consists of recycling electronic and robot disposals is adopted. Students are engaged in a real reverse engineering process and then challenged to find new possible applications and uses.

By adopting a pedagogical perspective, this paper introduces the design and implementation of a robot control system on a hardware platform based on a *Programmable Logic Controller* (PLC). In particular, the controlled robot is a *Sykerobot 600-5* manipulator with five degrees of freedom (DOFs) that was disposed of by the industry several years ago as electronic waste. Particular emphasis is placed on the pedagogical effectiveness of the proposed control architecture.

INTRODUCTION

Automation engineering education is a multidisciplinary field of study that involves different types of knowledge and skills. This educational field applies the discipline of mechanical systems, electronic systems, computers and control systems to the integration of product design and automated manufacturing processes. The Automation engineering program at the Faculty of Engineering and Natural Sciences and the Product and System Design program at the faculty of Maritime Technology and Operation, at Aalesund University College (AAUC), Norway, provide courses leading to Bachelor's and Master's degrees. These two study programs have several common topics concerning automation engineering subjects.

A common teaching strategy of these programs involves the ideas of Learning by Doing (LBD) (Nguyen & Graefe 2001),

the approaches of Problem Based Learning (PBL) (Albanese & Mitchell 1993) and the concepts of Active Learning (AL) (Martín et al. 2010). In fact, one of the most effective ways of teaching students how to perform a useful task consists of actively involving them and letting them do it. The LBD method is not a new instructional theory, it is exactly what it sounds like. Aristotle stated: *“One must learn by doing the thing, for though you think you know it, you have no certainty until you try”*. Similarly, Confucius declared: *“I hear and I forget. I see and I remember. I do and I understand”*. More recently, John Dewey became one of the strongest proponents of the LBD approach. In (Dewey 1997), Dewey argued: *“Education is not preparation for life, it is life itself”*.

At AAUC, during their study courses, students are involved with realistic problem settings and scenarios that reflect real application perspectives (Rekdalsbakken & Sanfilippo in press). Very often, students are divided into groups that stimulate their teamwork skills and critical thinking abilities. From a social point of view, group dynamics are also relevant. In order to prepare the students for their working life, the preferred method of putting groups together is randomly, with a random leader. This method is perceived as fair by the students. Moreover, normal working conditions are simulated in which the team members are usually unable to select their own team. In addition, this approach also establishes new social networks in the classroom. Our experience is that the students perform better when they know each other well. This probably has to do with the fact that they feel safer in the learning environment and are less afraid of possibly embarrassing situations. However, in generating random groups, an attempt is made to break up the existing frozen social ties, thereby forcing the students into new roles. As such, an industry-like social situation is created.

Moreover, our students are included in research projects and innovation activities in cooperation with real companies and industry partners. In such a view, the student laboratory has a central and challenging position as an open-space workplace where students can experience hands-on automation integration training under the supervision of both their professors and the partner company engineers. The networking between students and companies allows the students to gain deeper knowledge about industry demands and challenges. The industry also gets valuable information for their recruitment processes and learns about ongoing research projects at the university. In addition

to inspiring and motivating students in their studies, AAUC regularly organises several internal robotic competitions and workshop events. The best student projects are often selected to join national and international robotic contests.

A great effort is made at AAUC to provide the students with an adequate and effective automation integration experience without major capital investments. Moreover, the idea of recycling out-of-date electronic equipment and robots is promoted. Stressing the fact that after a robot has outlived its normal utility, its disposal becomes a challenge for the enterprises using it, students are challenged to find new possible applications and uses.

One of the most challenging robotics engineering tasks involves the integration of a robotic arm in material handling, assembly, and production processes. The knowledge and skills required for these kinds of tasks are purely mechatronic and therefore multidisciplinary. Emphasising the pedagogical prospective, this paper introduces the design and implementation of a robot control system on a hardware platform based on a *Programmable Logic Controller* (PLC) (Bolton 2009). The controlled arm is a *Sykerobot 600-5* manipulator with five DOFs that was disposed of by the industry several years ago as electronic waste. A master-slave architecture is set up with the controller acting as a master and the PLC as a slave. The paper analyses the drawbacks and the advantages related to the choice of standard PLCs in these kinds of applications, compared to the much more common choice of specialised hardware or industrial proprietary computers. Particular emphasis is placed on the pedagogical effectiveness of the proposed control architecture.

This paper is organised as follows. A review of the related research work is given in the second Section. In the third Section, we focus on the description of the system architecture. In the fourth Section, related results are discussed. Finally, conclusions and future works are outlined in the fifth Section.

RELATED RESEARCH WORK

AAUC has made a notable effort in order to limit the financially demanding cost of acquiring state-of-the-art manufacturing equipment. For instance, in (Liu et al. 2012), our research group presented a modular pentapedal walking robot that can be also used for pedagogical uses. Similarly, several university laboratories have followed different strategies.

One possibility consists of developing virtual laboratories and workspaces that can provide the students with an acceptable learning experience. In (Callaghan et al. 2008), for instance, the popular virtual world, *Second Life*, is used as a platform to create experiential based learning experiences in a 3-D immersive world for teaching computer hardware and electronic systems. In particular, a number of approaches to capturing, displaying and visualising real world data in such environment are implemented. The main goal of this virtual laboratory is to allow students to easily interact with a set of physical processes via the Internet. The students are able to run experiments, change control parameters, and analyse the results remotely. An additional feature of this virtual laboratory is its architecture, allowing for an easy integration of new processes for control experiments. In (Zhang et al. 2007), Zhang et. al. presented a kind of educational robotics

system based on the use of *LEGO* bricks and on a newly designed input/output interface. Using this system, students can program a robot through an iconic interface environment and a normal programming language such as Java or C according to their knowledge. During this process, they learn the sensorial technology and motor-control methods. At the same time, students can overview the process using a web-camera and can interrupt it in case of malfunctions. However, the advantages and benefits enjoyed by students that work in a real physical laboratory can hardly be replaced by any virtual counterparts.

To meet the need of providing the students with a physical experience without major capital investments, general purpose open-source developing platforms could be used as pedagogical tools. In (Sarik & Kymissis 2010), Sarik and Kymissis presented a lab kit platform based on an *Arduino* micro-controller board and open hardware that enables students to use low-cost, course specific hardware to complete lab exercises at home. This somehow represents an extension of the university laboratory and gives students the possibility of improving their learning experience. However, this approach does not provide the students with a real industry-like experience.

One possible way of providing students with a real industry-like experience consists of using PLC-based developing platforms. In (Chung 1998), Chung presented a cost-effective approach for the development of an integrated PC-PLC-Robot system for industrial engineering education. This work shows that even though many universities do not have the financial resources to acquire state-of-the-art manufacturing systems, they can still provide their students with an adequate and effective integration training with existing equipment. Our approach in this paper, follows the same idea, emphasising the effectiveness from an educational point of view.

SYSTEM ARCHITECTURE

The controlled robot is a *Sykerobot 600-5* manipulator with five DOFs. This robot was disposed of by one of our industry partners several years ago as electronic waste. Since this manipulator is obsolete, it is relatively hard for students to find any related works. This fact is particularly relevant from a pedagogical point of view because it forces students to get thoroughly exposed to the subject and involves them in a real reverse engineering process. Moreover, since the original controller cabinet of the robot is missing, each group of students need to develop its own control architecture. According to the teacher's experience, the most promising system solution developed by the students is presented in the following of this paper.

The system architecture is shown in Fig. 1. By using the *Modbus* protocol (Modbus 2004), a master-slave architecture is set up with the controller acting as a master and the PLC as a slave. The control software is fully developed on a commercial PLC system, using its standard programming tools and the multi-tasking features of its operating system. The input device is connected to the computer through the serial USB channel. In the next subsections, the different components of the system are described.

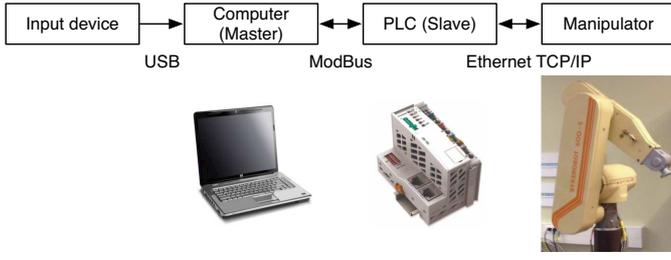


Fig. 1. The proposed control system architecture: a master-slave architecture with the controller acting as a master and the PLC as a slave.

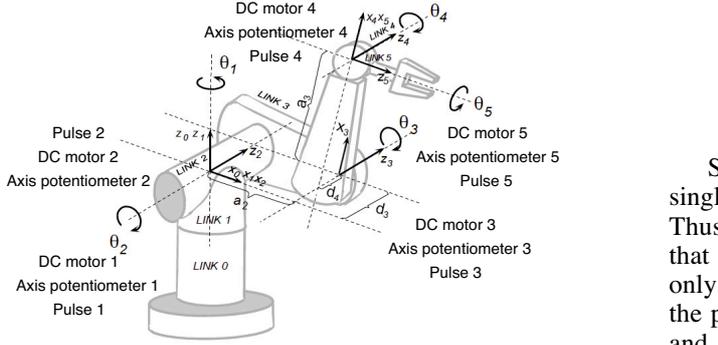


Fig. 2. The Sykerobot 600-5 manipulator with 5 DOFs.

The Control Algorithm

The kinematic sketch of the Sykerobot 600-5 is shown in Fig. 2. A good exercise for students consists of finding the kinematic model of the arm. Students learn about the use of geometric transformations, also called rigid transformations, to describe the movement of components in a mechanical system. These transformations simplify the derivation of the equations of motion, and are central to dynamic analyses.

According to the frame assignments in Fig. 2, the Denavit-Hartenberg (D-H) tables (Denavit 1955) of the Sykerobot 600-5 is shown in Table I. Substituting the DH parameters into the following general homogeneous transformation (HT) matrix,

$${}_{i-1}T_i = \begin{bmatrix} c\theta_i & -s\theta_i & 0 & a_{i-1} \\ s\theta_i c\alpha_{i-1} & c\theta_i c\alpha_{i-1} & -s\alpha_{i-1} & -s\alpha_{i-1}d_i \\ s\theta_i s\alpha_{i-1} & c\theta_i s\alpha_{i-1} & c\alpha_{i-1} & c\alpha_{i-1}d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

where s stands for \sin and c for \cos , the relative HT matrices for the manipulator can be obtained:

$${}^0_1T = \begin{bmatrix} c\theta_1 & -s\theta_1 & 0 & 0 \\ s\theta_1 & c\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (2)$$

$${}^1_2T = \begin{bmatrix} c\theta_2 & -s\theta_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -s\theta_2 & -c\theta_2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (3)$$

$${}^2_3T = \begin{bmatrix} c\theta_3 & -s\theta_3 & 0 & a_2 \\ s\theta_3 & c\theta_3 & 0 & 0 \\ 0 & 0 & 1 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

TABLE I. D-H TABLE OF THE Sykerobot 600-5, WHERE $a_2 = 0.33m$, $a_3 = 0.27m$, $d_3 = 0.20m$ AND $d_4 = 0.09m$

i	α_{i-1}	a_{i-1}	d_i	θ_i
1	0	0	0	θ_1
2	$-\frac{\pi}{2}$	0	0	θ_2
3	0	a_2	d_3	θ_3
4	0	a_3	$-d_4$	θ_4
5	$-\frac{\pi}{2}$	0	0	θ_5

$${}^3_4T = \begin{bmatrix} c\theta_4 & -s\theta_4 & 0 & a_3 \\ s\theta_4 & c\theta_4 & 0 & 0 \\ 0 & 0 & 1 & -d_4 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (5)$$

$${}^4_5T = \begin{bmatrix} c\theta_5 & -s\theta_5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -s\theta_5 & -c\theta_5 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (6)$$

Since the two joint axes of the arm's wrist intersect in a single point, it is useful to consider arm and wrist separately. Thus, the arm part is defined as the part of the manipulator that contributes to the position of the wrist, while the wrist only changes its orientation (the wrist itself does not affect the position). In this case, the arm part consists of links 0–3 and a part of the link 4. Since the wrist does not have any length parameters ($a_4 = d_5 = 0$), its relative HT matrix only has pure rotations. Consequently the HT matrix of the arm is:

$$T_A = {}^0_1T_2T_3T_{Screw}(e_1, \alpha_3, a_3)Trans(e_3, -d_4), \quad (7)$$

where $Screw(e_1, \alpha_3, a_3)$ represents the *Screw* of the reference frame $\{4\}$, while $Trans(e_3, -d_4)$ is the translation of the same reference frame along z_3 by $-d_4$. After multiplying the parts, we can get the forward kinematic (FK) equations. In detail, the arm rotation and position matrices, R_A and p_A respectively, are calculated:

$$R_A = \begin{bmatrix} c(\theta_2 + \theta_3)c\theta_1 & -s(\theta_2 + \theta_3)c(\theta_1) & -s\theta_1 \\ c(\theta_2 + \theta_3)s\theta_1 & -s(\theta_2 + \theta_3)s(\theta_1) & c\theta_1 \\ -s(\theta_2 + \theta_3) & -\cos(\theta_2 + \theta_3) & 0 \end{bmatrix}, \quad (8)$$

$$p_A = \begin{bmatrix} d_4s\theta_1 - d_3s\theta_1 - a_3(c\theta_1s\theta_2s\theta_3 - c\theta_1c\theta_2c\theta_3) + a_2c\theta_1c\theta_2 \\ d_3c\theta_1 - a_3(s\theta_1s\theta_2s\theta_3 - c\theta_2c\theta_3s\theta_1) - d_4c\theta_1 + a_2c\theta_2s\theta_1 \\ -a_3s(\theta_2 + \theta_3) - a_2s(\theta_2) \end{bmatrix}. \quad (9)$$

Up to this point, the forward position equations relating joint positions and end-effector positions and orientations have been derived. In the following, the velocity relationships, that relate the linear and angular velocities of the end-effector (or any other point on the manipulator) to the joint velocities are derived. Mathematically, the FK equations define a function between the space of Cartesian positions and orientations and the space of joint positions. The velocity relationships are then determined by the Jacobian of this function. The Jacobian is a matrix-valued function and can be thought of as the vector version of the ordinary derivative of a scalar function. The Jacobian matrix is one of the most important pieces of information in the analysis and control of robot motion.

The robot considered presents only spherical joints, therefore, the description of the angular velocity, ${}^{i+1}\omega_{i+1}$, of link $i+1$ can be obtained as:

$${}^{i+1}\omega_{i+1} = {}^i{}^{i+1}R^i\omega_i + \dot{\theta}_{i+1}{}^{i+1}\hat{z}_{i+1}, \quad (10)$$

where ${}^{i+1}R$ is the rotation matrix of frame $\{i\}$ with respect to $\{i+1\}$, ${}^i\omega_i$ is the angular velocity of frame $\{i\}$, $\dot{\theta}_{i+1}$ is the angular velocity of joint $i+1$ and ${}^{i+1}\hat{z}_{i+1}$ is the unit vector of frame $\{i+1\}$. Similarly, the corresponding relationship for the linear velocity, ${}^{i+1}v_{i+1}$, of link $i+1$ is given by:

$${}^{i+1}v_{i+1} = {}^{i+1}R({}^i v_i + {}^i\omega_i \times {}^i P_{i+1}), \quad (11)$$

where ${}^i v_i$ is the linear velocity of frame $\{i\}$ and ${}^i P_{i+1}$ is the position of frame $\{i+1\}$ respect to $\{i\}$. Applying these equations successively from link to link, we can compute ${}^N\omega_N$ and ${}^N v_N$, the rotational and linear velocity of the last link. For the considered arm, we get:

$${}^1\omega_1 = \begin{bmatrix} 0 \\ 0 \\ \dot{\theta}_1 \end{bmatrix}, {}^1v_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad (12)$$

$${}^2\omega_2 = \begin{bmatrix} -\dot{\theta}_1 s\theta_2 \\ -\dot{\theta}_1 c\theta_2 \\ \dot{\theta}_2 \end{bmatrix}, {}^2v_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad (13)$$

$${}^3\omega_3 = \begin{bmatrix} -\dot{\theta}_1 c\theta_2 s\theta_3 - \dot{\theta}_1 c\theta_3 s\theta_2 \\ \dot{\theta}_1 s\theta_2 s\theta_3 - \dot{\theta}_1 c\theta_2 c\theta_3 \\ \dot{\theta}_2 + \dot{\theta}_3 \end{bmatrix}, \quad (14)$$

$${}^3v_3 = \begin{bmatrix} s\theta_3(a_2\dot{\theta}_2 + d_3\dot{\theta}_1 s\theta_2) - d_3\dot{\theta}_1 c\theta_2 c\theta_3 \\ c\theta_3(a_2\dot{\theta}_2 + d_3\dot{\theta}_1 s\theta_2) + d_3\dot{\theta}_1 c\theta_2 s\theta_3 \\ a_2\dot{\theta}_1 c\theta_2 \end{bmatrix}, \quad (15)$$

$${}^4\omega_4 = \begin{bmatrix} -\dot{\theta}_1 c\theta_2 s\theta_3 - \dot{\theta}_1 c\theta_3 s\theta_2 \\ \dot{\theta}_1 s\theta_2 s\theta_3 - \dot{\theta}_1 c\theta_2 c\theta_3 \\ \dot{\theta}_2 + \dot{\theta}_3 + \dot{\theta}_4 \end{bmatrix},$$

$${}^4v_4 = \begin{bmatrix} a_2\dot{\theta}_2 s\theta_3 - d_3\dot{\theta}_1 c(\theta_2 + \theta_3) + d_4\dot{\theta}_1 c(\theta_2 + \theta_3) \\ a_3\dot{\theta}_2 + a_3\dot{\theta}_3 + a_2\dot{\theta}_2 c\theta_3 + d_3\dot{\theta}_1 s(\theta_2 + \theta_3) - d_4\dot{\theta}_1 s(\theta_2 + \theta_3) \\ a_2\dot{\theta}_1 c\theta_2 + a_3\dot{\theta}_1 c(\theta_2 + \theta_3) \end{bmatrix}.$$

To find these velocities with respect to the non-moving base frame, they can be rotated by using the rotation matrix R_A :

$${}^0v = R_A {}^4v_4 = \begin{bmatrix} 0 \\ {}^4v_{1,1} \\ {}^4v_{2,1} \\ {}^4v_{3,1} \end{bmatrix},$$

$${}^0v_{1,1} = d_4\dot{\theta}_1 c\theta_1 - d_3\dot{\theta}_1 c\theta_1 - a_2\dot{\theta}_1 c\theta_2 s\theta_1 - a_2\dot{\theta}_2 c\theta_1 s\theta_2 - a_3\dot{\theta}_1 c\theta_2 c\theta_3 s\theta_1 - a_3\dot{\theta}_2 c\theta_1 c\theta_2 s\theta_3 - a_3\dot{\theta}_2 c\theta_1 c\theta_3 s\theta_2 - a_3\dot{\theta}_3 c\theta_1 c\theta_2 s\theta_3 - a_3\dot{\theta}_3 c\theta_1 c\theta_3 s\theta_2 + a_3\dot{\theta}_1 s\theta_1 s\theta_2 s\theta_3, \quad (16)$$

$${}^0v_{2,1} = d_4\dot{\theta}_1 s\theta_1 - d_3\dot{\theta}_1 s\theta_1 - a_2\dot{\theta}_2 s\theta_1 s\theta_2 + a_2\dot{\theta}_1 c\theta_1 c\theta_2 + a_3\dot{\theta}_1 c\theta_1 c\theta_2 c\theta_3 - a_3\dot{\theta}_1 c\theta_1 s\theta_2 s\theta_3 - a_3\dot{\theta}_2 c\theta_2 s\theta_1 s\theta_3 - a_3\dot{\theta}_2 c\theta_3 s\theta_1 s\theta_2 - a_3\dot{\theta}_3 c\theta_2 s\theta_1 s\theta_3 - a_3\dot{\theta}_3 c\theta_3 s\theta_1 s\theta_2,$$

$${}^0v_{3,1} = -a_2\dot{\theta}_2 c\theta_2 - a_3\dot{\theta}_2 c(\theta_2 + \theta_3) - a_3\dot{\theta}_3 c(\theta_2 + \theta_3).$$

As such, the time derivative of the kinematics equations yields the *Jacobian* matrix of the arm, which relates the joint

rates to the linear and angular velocity:

$$J = \begin{bmatrix} J_{1,1} & J_{1,2} & J_{1,3} \\ J_{2,1} & J_{2,2} & J_{2,3} \\ J_{3,1} & J_{3,2} & J_{3,3} \end{bmatrix},$$

$$J_{1,1} = d_4 c\theta_1 - d_3 c\theta_1 - a_2 c\theta_2 s\theta_1 - a_3 c\theta_2 c\theta_3 s\theta_1 + a_3 s\theta_1 s\theta_2 s\theta_3,$$

$$J_{1,2} = -a_2 c\theta_1 s\theta_2 - a_3 c\theta_1 c\theta_2 s\theta_3 - a_3 c\theta_1 c\theta_3 s\theta_2,$$

$$J_{1,3} = -a_3 c\theta_1 c\theta_2 s\theta_3 - a_3 c\theta_1 c\theta_3 s\theta_2,$$

$$J_{2,1} = d_4 s\theta_1 - d_3 s\theta_1 + a_2 c\theta_1 c\theta_2 + a_3 c\theta_1 c\theta_2 c\theta_3 - a_3 c\theta_1 s\theta_2 s\theta_3,$$

$$J_{2,2} = -a_2 s\theta_1 s\theta_2 - a_3 c\theta_2 s\theta_1 s\theta_3 - a_3 c\theta_3 s\theta_1 s\theta_2,$$

$$J_{2,3} = -a_3 c\theta_2 s\theta_1 s\theta_3 - a_3 c\theta_3 s\theta_1 s\theta_2,$$

$$J_{3,1} = 0,$$

$$J_{3,2} = -a_2 c\theta_2 - a_3 c(\theta_2 + \theta_3),$$

$$J_{3,3} = -a_3 c(\theta_2 + \theta_3).$$

PLC

Since the discarded robot is missing the controller cabinet, students are encouraged to develop their own control system on a PLC architecture. A PLC is a type of digital computer that is generally used in automation for electro-mechanical processes, typically for industrial use. A PLC can be controlled by a simulation program designed on a computer and it is equipped with a set of Digital Inputs (DI), Digital Outputs (DO), Analog Inputs (AI) and Analog Outputs (AO) or Pulse-width modulation (PWM) outputs. This kind of I/O interface is typically conform to strict industrial quality standards with protected inputs (often galvanically separated from the PLC by optocouplers) and outputs. The operating range is commonly at 24V or 4-20mA signal levels. These characteristics are relevant from a didactic point of view, giving the students the opportunity of experience a typical industrial architecture setup. Moreover, a PLC can be logically programmed in different forms, such as a ladder diagram, a structural text and a functional block diagram and stored in memory. These different programming possibilities give students the chance to learn different programming techniques and approaches. A PLC is an example of a hard real-time system since output results must be produced in response to input conditions within a limited time, otherwise an unintended operation will result. These strict requirements force students to design and implement reliable and efficient software.

Control Architecture

The control architecture is shown more in detail in Fig. 3. The *Sykerobot 600-5* manipulator has five axes which are driven by DC motors (24Vdc). Each DC motor is connected to a gear mechanism that provides feedback to two position sensors, a potentiometer and a quadrature pulse transmitter, as shown in Fig. 4 for one of the joint. From the gear box of the DC motor, the output of the motor is delivered via servo spline to the servo arm. The potentiometer's changes in position correspond with the current position of the motor. Therefore, the change in resistance produces an equivalent change in voltage from the potentiometer. The quadrature outputs (A and B signals that are separated by 90 electrical degrees) are feed into suitable decoders/counters that are able to detect direction reversal due to the quadrature feature.

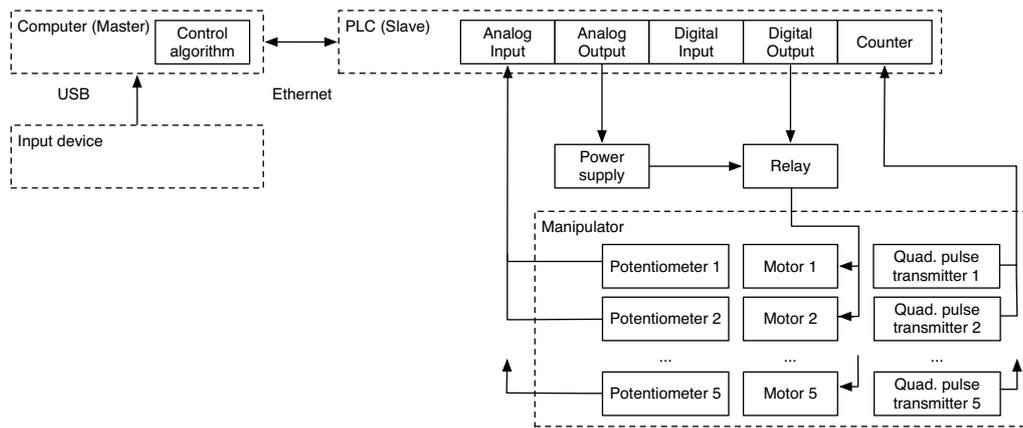


Fig. 3. The proposed control system architecture.



Fig. 4. A detailed photo of the potentiometer and of the quadrature pulse transmitter from one of the manipulator joint.

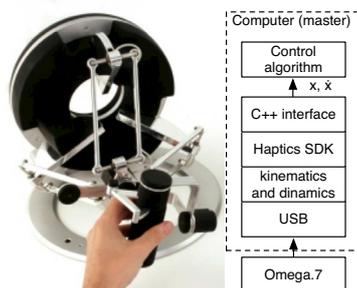


Fig. 5. The *omega.7* haptic device from *Force Dimension* that is used as an input device is shown on the left side, while the corresponding data work-flow is outlined on the right side.

Since the control cabinet is no longer available, the motors must be interfaced to some kind of motor controller. In order to avoid buying costly *H bridge* circuits, a programmable power supply board is used. This board can be remotely controlled through a 0-5V signal from the PLC's AO. Besides, the motor revolution direction (clockwise or counterclockwise) is controlled by reversing the polarity with the use of relays.

Input Device

In this case study, a commercial haptic device, the *omega.7* from *Force Dimension*, is used as the input for the system. This device is shown on the left side of Fig. 5 and it is considered state-of-the-art in this field. This choice is justified by the AAUC's goal of providing the students with some of

the newest technologies, as well as by the adopted recycling policy. Furthermore, from a pedagogical point of view, the integration of new technologically advanced devices with out-of-date disposed electronics engages students in challenging tasks. The integration of the *omega.7* is realised by using the *Haptics SDK* provided by *Force Dimension*, as shown on the right side of Fig. 5. The position is read by using a C++ interface and used as the input for the control algorithm.

The *omega.7* is a seven DOFs haptic interface with high precision active grasping abilities and orientation sensing. Finely tuned to display perfect gravity compensation, its force-feedback gripper offers extraordinary haptic features, enabling instinctive interaction with complex haptic applications. Since this particular input device presents a higher number of DOFs compared to the controlled robot, the students are challenged to find a mapping approach. A quite interesting solution implemented by the students consists of using the first three DOFs of the *omega.7* to specify the desired position, while the next three DOFs are utilised to set the end-effector orientation. Finally, the seventh DOF is reserved to control a possible tool to be mounted on the manipulator tip.

It should be noted that thanks to the modularity of the proposed system architecture, a different input device can be also used without affecting the effectiveness of the proposed method.

RESULTS

During this learning experience, students have the chance to involve themselves into realistic challenges in the design and implementation of complex systems and to integrate the knowledge and skills gained during their courses. The LBD, PBL, and AL approaches all share a closed loop learning process where the learners get immediate and objective feedback on their progress towards solving the problem at hand. This assimilation process is illustrated in Fig. 6.

Even though our students are undergraduate students, they experience the same benefits that Papert observed in high school students (Papert 1980). He emphasised that learning takes place easily when knowledge fits into the students' learning model: "Anything is easy if you can assimilate it to your collection of models. If you can't, anything can be painfully difficult".

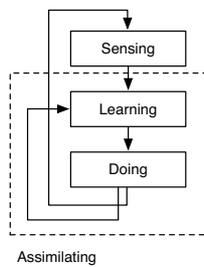


Fig. 6. The effective learning paradigm.

CONCLUSION AND FUTURE WORK

In this work, a combination of the LBD, PBL, and AL approaches is applied to a particular case study: recycling a discarded robotic arm for automation engineering education. According to the feedback received by our students, this experience has shown positive results and improvements on both a learning and a social level. When giving the students the possibility of doing and applying theoretical knowledge on practical experiences, the assimilation process is faster and the social climate of the class improves.

The proposed approach enables students to gain practical knowledge of the integration of different engineering fields, including mechanics, programmable logic controllers, BUS systems, kinematics and control systems. A team learning strategy is proposed and support to hands-on activities in an open-space laboratory is provided. One of the most important learning gains for students consists of getting familiar with different engineering fields by working through a scenario that simulate some challenging industrial tasks and conditions.

According to the author's experience, the involvement of students with triggering and inspiring tasks results in their acquiring new skills and knowledge at higher levels of learning, including analysis, synthesis and evaluation.

As future work, this same learning approach can be applied to new groups of students in order to certify the effectiveness by constantly monitoring them with a set of targeting questions and surveys.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of the students Bjarne André Humlen, Anders Bakken Kobbevik, Simon Langlo, Daniel Nedregård.

REFERENCES

- Albanese, M. A. & Mitchell, S. (1993), 'Problem-based learning: A review of literature on its outcomes and implementation issues', *Academic medicine* **68**(1), 52–81.
- Bolton, W. (2009), *Programmable logic controllers*, Access Online via Elsevier.
- Callaghan, M., Harkin, J., Scibilia, G., Sanfilippo, F., McCusker, K. & Wilson, S. (2008), Experiential based learning in 3d virtual worlds: Visualization and data integration in second life, in 'Remote Engineering & Virtual Instrumentation, (REV 2008)', REV.
- Chung, C. A. (1998), 'A cost-effective approach for the development of an integrated pc-plc-robot system for industrial

- engineering education', *IEEE Transactions on Education* **41**(4), 306–310.
- Denavit, J. (1955), 'A kinematic notation for lower-pair mechanisms based on matrices.', *Trans. of the ASME. Journal of Applied Mechanics* **22**, 215–221.
- Dewey, J. (1997), *How we think*, Courier Dover Publications.
- Liu, C., Sanfilippo, F., Zhang, H., Hildre, H. P., Liu, C. & Bi, S. (2012), Locomotion analysis of a modular pentapedal walking robot, in 'Proceeding of the European Conference on Modeling and Simulation (ECMS)', pp. 441–447.
- Martín, E., Lázaro, C. & Hernan-Losada, I. (2010), Active learning in telecommunication engineering: A case study, in 'Education Engineering (EDUCON)', IEEE, pp. 1555–1562.
- Modbus, I. (2004), 'Modbus messaging on tcp', *IP Implementation Guide v1.0a*, North Grafton, Massachusetts (www.modbus.org/specs.php).
- Nguyen, M.-C. & Graefe, V. (2001), Learning by doing-an approach to robotic skill acquisition, in 'SICE 2001. Proceedings of the 40th SICE Annual Conference. International Session Papers', IEEE, pp. 226–229.
- Papert, S. (1980), *Mindstorms: Children, computers, and powerful ideas*, Basic Books, Inc.
- Rekdalsbakken, W. & Sanfilippo, F. (in press), Enhancing undergraduate research and learning methods on real-time processes by cooperating with maritime industries, in 'Proceeding of the European Conference on Modeling and Simulation (ECMS 2014)'.
- Sarik, J. & Kymissis, I. (2010), Lab kits using the arduino prototyping platform, in 'Frontiers in Education Conference (FIE), 2010 IEEE', IEEE, pp. T3C–1.
- Zhang, H., Zheng, W., Chen, S., Zhang, J., Wang, W. & Zong, G. (2007), Flexible educational robotic system for a practical course, in 'Integration Technology, 2007. ICIT'07. IEEE International Conference on', IEEE, pp. 691–696.

AUTHOR BIOGRAPHIES

FILIPPO SANFILIPPO is a PhD candidate in Engineering Cybernetics at the Norwegian University of Science and Technology, and a research assistant at the Department of Maritime Technology and Operations, Aalesund University College, Norway. He obtained his Master's Degree in Computer Engineering at University of Siena, Italy.
Email: fisa@hials.no

OTTAR L. OSEN received his M.Sc. in Cybernetics in 1992 at the Norwegian University of Science and Technology and holds the position of Assistant Professor at Aalesund University College, Norway. He also holds the position of Head of R&D at ICD Software in Aalesund, Norway
Email: oo@hials.no

SALEH ALALIYAT was born in Jenin, Palestine. He is currently working as a PhD candidate at Aalesund University College, Norway. He received his Masters degree in Media Technology from Gjvik University College in Norway.
Email: saal@hials.no

MODELLING AND SIMULATION OF AN OFFSHORE HYDRAULIC CRANE

Yingguang Chu, Vilmar Æsøy and Houxiang Zhang
Department of Maritime Technology and Operations
Aalesund University College
PO Box 1517, N-6025 Ålesund, Norway
Email: yich@hials.no
ve@hials.no
hozh@hials.no

Øyvind Bunes
Deck Machinery Seismic & Subsea
Rolls-Royce Marine
PO Box 193, N-6069 Hareid, Norway
Email: Oyvind.Bunes@Rolls-Royce.com

KEYWORDS

Bond Graph, offshore crane, hydraulic system, modeling and simulation

ABSTRACT

This paper presents a modeling approach based on Bond Graph (BG) method for offshore hydraulic crane focusing on its hydraulic system characteristics. A hydraulic library is built in the modeling software tool 20-sim using BG elements. The hydraulic submodels are designed according to one specific type of offshore crane, however, they can be easily modified and reused for other similar systems. BG method is a modelling technique for modeling of complex system by describing the energy flow inside the physical system. One of the main benefits of modeling using BG for the hydraulic system is the model provide interfaces to systems of other domains, for example, cooling system, mechanical model, control unit, etc. In this paper it is shown how an integrated BG model of the hydraulic system for a knuckle boom crane is derived and used for simulation. The simulation results proved the validation and effectiveness of the presented modeling approach for simulation of multi-domain systems.

INTRODUCTION

Cranes are found onboard almost all kinds of vessels and platforms for handling personnel and cargo. Cranes onboard vessels and platforms handling goods between the quayside and vessel or between vessels are normally referred to as offshore cranes. Cranes that are used for handling submerged loads as well e.g. launch and recovery of submersibles or installation of subsea hardware, are normally referred to as subsea cranes. Compare to land based cranes with a solid fixed base, offshore and subsea cranes are subject to significant dynamic forces from the resulting payload sway directly or indirectly caused by the vessel motion. As field testing in offshore industry is expensive and time consuming to carry out and constrained by many factors such as weather condition and vessel availability,

modeling and simulation become a crucial part for product design, testing and analysis.

On one hand, offshore cranes are mostly hydraulic actuated due to the consideration for stable performance and safety redundancy. On the other hand, it is rather delicate to model and control hydraulic systems because of the complex dynamic behavior and nonlinear aspect of fluid energy transfer. Many studies on hydraulic system modeling dedicated to one or several specific components. There are many software tools available for modelling and simulation of hydraulic systems. Modelling tools used in former researches include SimHydraulic from MathWorks (Vêchet and Krejsa 2009), Easy5 from MSC (Li et al. 2011), SimulationX from ITI (En et al. 2013), 20-sim from Controllab (Aridhi et al. 2013), etc. These programs provide standard libraries for hydraulic components which can be parameterized and modified to certain levels.

The generalized models are not designed for a specific system which means they might be over-complicated thus compromise the simulation efficiency. It is possible, to a certain level, to create new specific models for components that are not included in these libraries from these software tools, but that's not always the best way. Take 20-sim as example, a hydraulic library is developed according to the Modelica hydraulic library. The library doesn't include all the valves in a crane system. Instead of using BG elements, the models are written in a way which is difficult for the users to understand and edit. In this paper we present a modeling approach for offshore hydraulic crane system based on BG method. The submodels are created from scratch using basic BG elements and are completely open for editing as detailed as necessary depending on the simulation purpose. Another reason of choosing 20-sim as the modelling tool is using BG method complex systems, e.g. an offshore hydraulic crane, involving multiple energy domains can be modelled and integrated.

The rest of the paper starts with introducing the basics of the BG method and the hydraulic system of the

kunckle boom crane. Then, the modelling of the main components using BG is described and the results from the simulation of the model are presented. Finally, the conclusion and future work is discussed.

BOND GRAPH METHOD

BG method as a general approach for modeling interacting systems is based on identifying the energetic structure in a system. A system can be decomposed into a few basic physical properties depending on what is going to be studied, and then the system can be described by interrelated idealized elements. The energy or power interaction between two elements is called a “power bond” represented by a half arrow. Another type of bond called “signal bond” represented by a full arrow indicates a signal flow at negligible power. A power bond is defined by two variables with generalized names of “effort” and “flow”, of which the product is power. Table 1 lists a number of energy domains and their corresponding power variables.

Table 1 Common used BG energy domains

Energy Domain	Effort (e)		Flow (f)	
	Name	Unit	Name	Unit
Mechanical translation	Force	N	Linear velocity	m/s
Mechanical rotation	Torque	Nm	Angular velocity	rad/s
Electrical	Voltage	V	Current	A
Hydraulic	Pressure	Pa	Volume flow	m ³ /s
Thermal	Temperature	K	Entropy flow	W/°C
Magnetic	Magneto-motive force	A	Flux rate	Wb/s
Chemical	Chemical potential	J/mol	Reaction rate	Mol/s

Roughly speaking, the basic elements account for energy supply based on supply of effort and flow (Se-element and Sf-element), potential and kinetic energy storage (C-element, I-element), energy dissipation (R-element) and energy transform (TF-element) or conversion (GY-element). In addition to the basic elements describing the boundary components, the interconnection in between two elements is described using an ideal 1-junction or 0-junction element, which neither store nor dissipate the energy. In brief, a 1-junction has equal flow on all bonds adjoining and the sum of efforts equals to zero, while a 0-junction is just the opposite: the effort is the same and the sum of flow is zero. The essence of defining an element is to establish the relation of the energy variables. Below Figure 1 so-called tetrahedron of state, illustrates the basic 1-port elements relating the energy variables (Pedersen and Engja 2008).

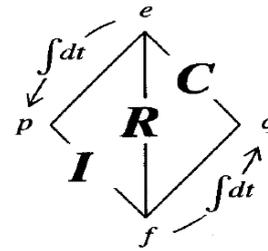


Figure 1: Tetrahedron of state for basic 1-port elements

OFFSHORE CRANE HYDRAULIC SYSTEM

The hydraulic system of a common offshore knuckle boom crane is studied in this paper. The crane consists of three joints actuated by a hydraulic motor and two hydraulic cylinders (Figure 2).

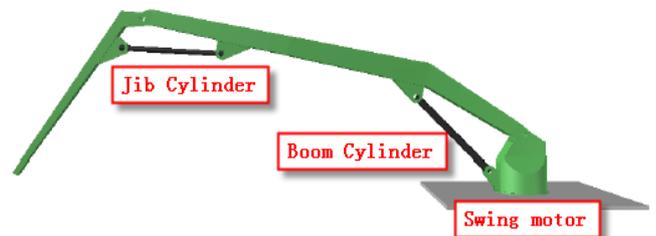


Figure 2: Offshore hydraulic knuckle boom crane

When considering the complexity of the model, it is vital that the simulation can be done in real time. Thus the hydraulic system schematic is simplified to include only the main components at a level corresponding to the characteristics that shall be studied (Figure 2). The main components of the crane hydraulic system include a Hydraulic Power Unit (HPU), pipelines, valves (compensator, 4/3proportional direction valve, load control valve), cylinders, and motors.

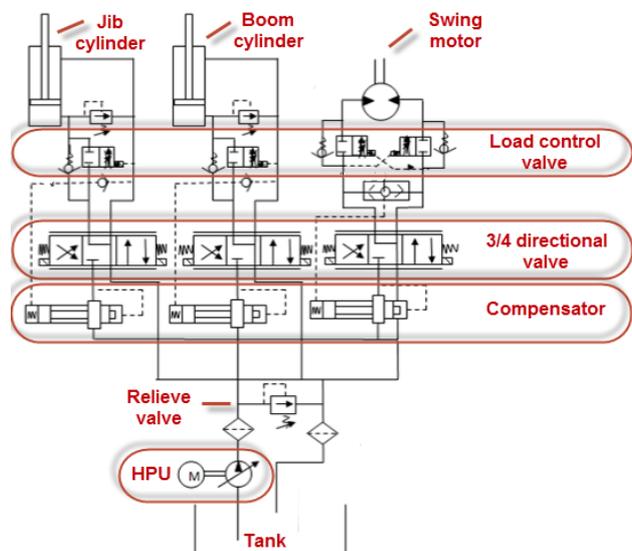


Figure 3: Hydraulic system schematic

BOND GRAPH MODELING OF CRANE HYDRAULIC SYSTEM

After identified the main components of the hydraulic system, in this chapter modeling of these components using BG elements is described. The hydraulic submodels are created based on the basic principles of fluid dynamics (ASSOFLUID 2007). To reduce the complexity of the overall model, the model of each component is also simplified. Fluid inertia and flexibility are dominant in the pipeline and cylinder chambers, thus neglected in the other components. As mentioned, BG method is modelling approach by describing the energy flow of the system. In the hydraulic domain, the key principle is to establish the connection of pressure and flow through the system.

HPU (pump)

The HPU of the crane mainly consist of a pressure compensated pump, which maintains a preset pressure at its outlet by adjusting its delivery flow in accordance with the pressure at any given time. If the system pressure is less than the pressure set point, the pump outputs its flow proportional to the pressure deviation. In the BG method a pump is modelled as a flow source element (Sf-element). The Sf-element has one output power port associated with the pump outlet. The effort and flow relationship is given by the following equations:

$$Q_{\max} = \frac{v}{2\pi} \omega \quad (1)$$

$$f = \frac{P_{\text{set}} - e}{\Delta P} Q_{\max} \quad (2)$$

Where v is the displacement of the pump, ω is the pump rotational speed, P_{set} is the pump pressure set point, ΔP pressure deviation from the set point that required giving full pump flow.

Pipe

The pipe submodel (Figure 4) describes segmental hydraulic pipelines with circular cross sections. The submodel accounts for friction loss along the pipe, fluid inertia and compressibility. The ControlVolume C-element is inserted in between pipe segments or other components as many as needed to avoid causality error and account for fluid flexibility in the areas that are not considered as pipes, i.e. with negligible inertia and frictional effects. The pipe sub-model has one input and one output power port associated with the physical inlet and outlet of the pipe segment.

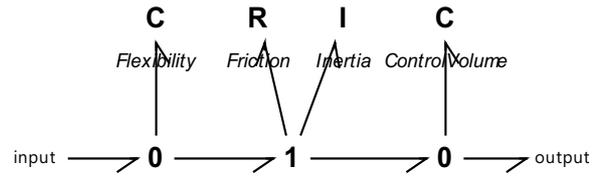


Figure 4: BG model of segmental hydraulic pipe

The Bond Graph elements are written by the following formulas:

Friction (R-element):

$$e = \frac{1}{2} k \frac{\rho l}{d} v^2 \quad (3)$$

Where the friction factor $k = \frac{64}{\text{Re}}$ for laminar flow given

by the Darcy–Weisbach equation, and $k = \frac{0.316}{\text{Re}^{0.25}}$ for

turbulence flow given by the Blasius equation. The Reynolds number is calculated by $\text{Re} = \frac{d}{Av} f$, ρ is the fluid density, v is the fluid viscosity, d is the pipe diameter, l is the pipe length.

Inertia (I-element):

$$\int e dt = \frac{\rho l}{A} f \quad (4)$$

Where ρ is the fluid density, l is the pipe segment length, A is the section area of the pipe.

Compressibility (C-element):

$$e = \frac{B}{Al} \int f dt \quad (5)$$

Where B is the fluid bulk modulus, A is the section area of the pipe, and l is the pipe segment length.

ControlVolume (C-element):

$$e = \frac{B}{V} \int f dt \quad (6)$$

Where B is the fluid bulk modulus, V is the volume of the fluid.

Valves

In the crane's hydraulic system, four types of valves are modeled; a relief valve, a compensator, a directional control valve and a load control valve. A valve submodel is described as a restriction nozzle which causes a pressure drop in the direction of flow. The relationship between the pressure and flow through the valve follows the general equation:

$$\dot{V} = c_d A \sqrt{\frac{2}{\rho} \Delta P} \quad (7)$$

Where \dot{V} is the flow rate, c_d is the discharging coefficient, A is the valve orifice area, ρ is the fluid density and ΔP is the pressure drop across the valve.

Compensator (MR-element):

The compensator submodel represents a flow control valve which maintains a certain pressure differential over a hydraulic valve to minimize the influence of pressure variation on a flow rate passing through that valve. The compensator model has one input power port associated with the valve inlet, and two input signal ports associated with the pressure on both sides of the compensated valve. The valve opening area is proportional to pressure differential:

$$A = \frac{f_{dp}}{\sqrt{\frac{2dp}{\rho}}} \quad (8)$$

Where is f_{dp} the flow rate at a pressure drop of dp . The value can be read from the flow chart in the valve factsheet.

The flow rate is then calculated by:

$$f = \frac{P_{set} - \Delta P}{P_{set}} A \sqrt{\frac{2e}{\rho}} \quad (9)$$

Where P_{set} is the pressure set point over the compensated component, ΔP is the pressure differential over the compensated component, A is the valve open area. ρ is the fluid density.

4/3proportional direction valve (R-elements):

The 4/3proportional direction valve submodel (Figure 5) represents a continuous 3-way-4-port directional valve. The fluid from the compensator is distributed between two output ports A and B and return to the output port Tank, varied by the slide position.

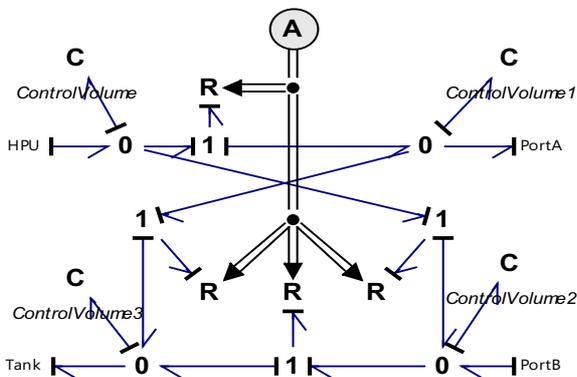


Figure 5: BG model of 4/3 proportional valve

The valve R-element is written based on the general equation:

$$f = A \sqrt{\frac{2e}{\rho}} \quad (9)$$

Where the opening area A is signaled by the valve slide position which is controlled via an external controller, for example a joystick or a keyboard.

Load control valve (MR-element):

The load control valve submodel consists of a pressure control valve and a check valve that can retard actuator's movement when with overrunning loads. The submodel has one input power port and one input signal port associated with the pressure signal at the load side. The valve flow rate is calculated based on Equation (9). The valve is signaled by the pressure at the load side. Flow is free in one direction while proportional to the load side pressure in the other direction.

Cylinder

The cylinder submodel (Figure 6) represents the hydraulic cylinder which converts hydraulic energy into mechanical energy in the form of translational motion. Hydraulic fluid pumped under pressure into one of the two cylinder chambers forces the piston to move and exert force on the cylinder rod. The cylinder can transfer force and motion in both directions. The cylinder submodel has one input power port and one output power port associated with the cylinder A and B port. The cylinder submodel has a second output power port associated with the cylinder output force.

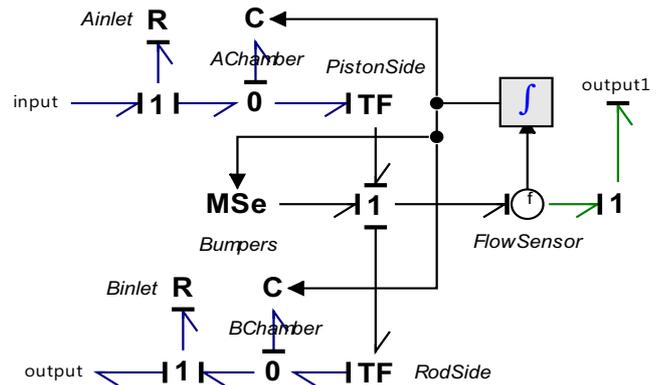


Figure 6: BG model of hydraulic cylinder

The cylinder inlets are modeled as restriction nozzles represented by R-elements. The relationship of power variables is given by Equation (9). The cylinder chambers are described as capacitors represented by C-elements. The power variables are given by Equation (6) where the volume is calculated by the cylinder piston position. The transformer elements (TF-element)

transform energy between hydraulic and mechanical domain. The power variables are given by the following equations:

$$e_1 = \frac{e_2}{A} \quad (10)$$

$$f_2 = \frac{f_1}{A} \quad (11)$$

Where e_1 is pressure, e_2 is force, A is the fluid contact area to the cylinder, f_1 is the flow rate, f_2 is the cylinder speed.

The cylinder model also includes limitations presented by an MSe-element. The cylinder end positions, fully retracted and fully extended, are modeled as a spring-damper systems which allows a certain deflection. The effort and flow relationship is given by:

$$e = -k\Delta x - cf \quad (12)$$

Where k is the stiffness, Δx is the deflection of the bumpers and c is the damping factor of the bumpers.

Motor

The motor submodel (Figure 7) represents a fixed-displacement hydraulic motor which converts hydraulic energy to mechanical energy in the form of rotational motion. The motor can transfer torque and rotation in both directions. The submodel has one input power port and one output power port associated with the motor inlet and outlet. The submodel has a second output power port associated with the rotational shaft. As this is a high speed motor, a reduction gear is required to transform high speed and low torque to low speed and high torque.

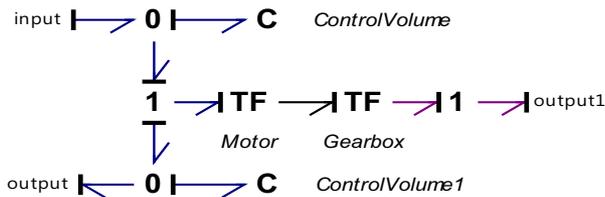


Figure 7: BG model of hydraulic motor and gearbox

The relationship of the power variables are given by the following formulas:

Motor (TF-element):

$$e_1 = \frac{v}{2\pi} e_2 \quad (13)$$

$$f_2 = \frac{v}{2\pi} f_1 \quad (14)$$

Where v is the displacement of the motor, e_1 is hydraulic pressure at motor, e_2 is output torque of motor, f_1 is flow through motor, and f_2 is motor rotational speed.

Gearbox (TF-element):

$$e_1 = \frac{1}{n} e_2 \quad (15)$$

$$f_2 = \frac{1}{n} f_1 \quad (16)$$

Where n is gear ratio.

Tank

The Tank submodel is modeled as an open effort source (Se-element) with one atmosphere pressure, i.e. 100000Pa.

$$e = 100000 \quad (19)$$

With all the BG models built for the components, a complete circuit of the hydraulic system can be assembled. Figure 8 (Page 6) shows the BG model for the boom cylinder circuit of the crane. The input signal is sent to the directional valve for controlling the slide position for flow distribution.

SIMULATION RESULTS

Similarly, the hydraulic models for the jib cylinder and slewing motor circuit can be created. Due to the size limit of the paper they are not shown but grouped in the integrated model (Figure 9). In this model the crane is controlled via a joystick and the crane body is represented simply by some mass and inertia elements. A more explicit model of the crane body can be developed using BG as well and connected to the hydraulic model.

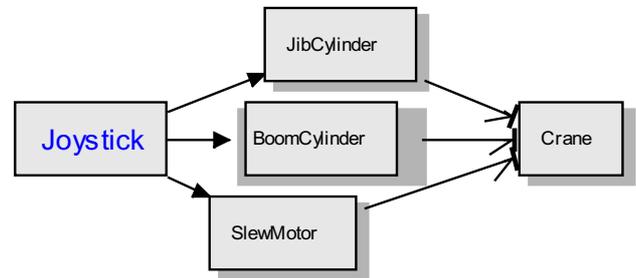


Figure 9: Integrated model of crane hydraulic system

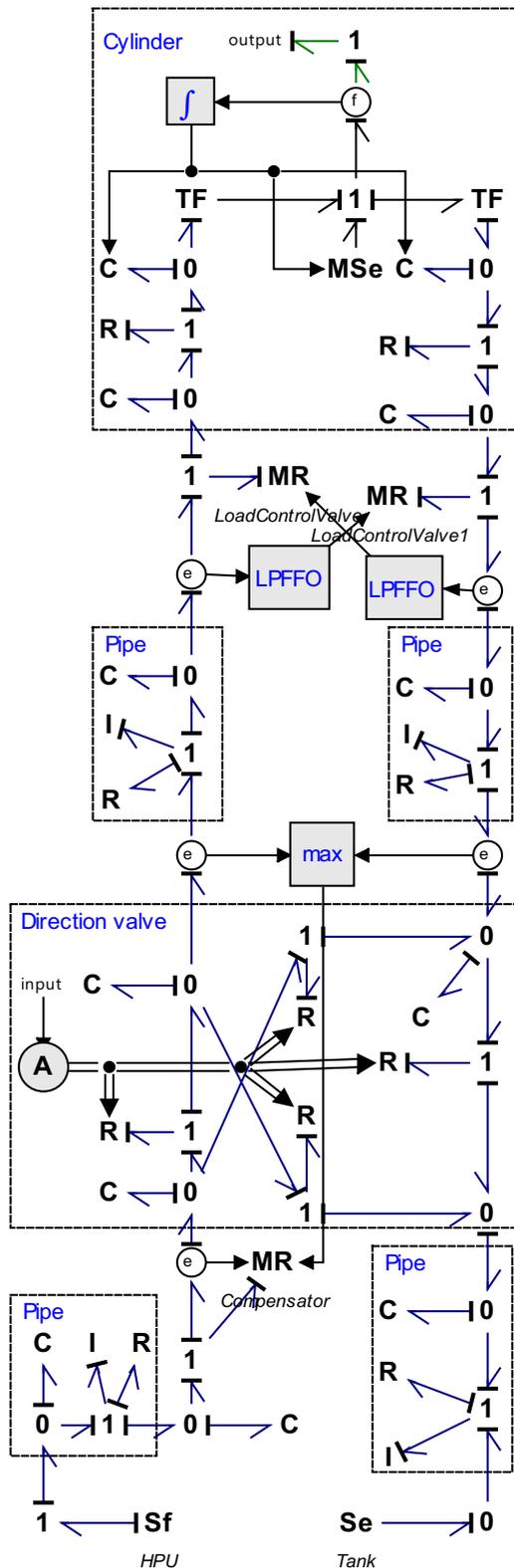


Figure 8: BG model of the crane boom cylinder circuit

All the variables in the model can be plotted during a real time simulation run. The following figures show the plotting of the boom cylinder circuit when the cylinder being extended and retracted. Figure 9 shows the boom cylinder position, which starts extending from its

minimum position of 0.1m at 5s until fully extended to 0.9m at around 25s. Then the boom cylinder is retracted back from 25s till reaching its fully retracted position at 48s. Correspondingly, Figure 10 shows the pressures in the boom cylinder chambers. The pressure starts increase from 5s and when the cylinder is fully extended, the pressure at A chamber continues accumulating until reaching the maximum pressure, while the pressure at B chamber drop to zero. From 25s, when the cylinder is being retracted, the pressure increases in B chamber and decreases in A chamber until fully retracted at 48s. Figure 11 shows the flow distribution through the directional valve.

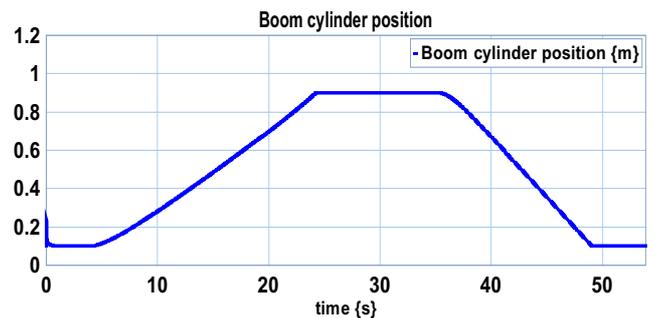


Figure 9: Boom cylinder position

The boom cylinder stroke is defined at 0.8m, with a minimum position at 0.1m to one end of the cylinder and 0.9m to the other.

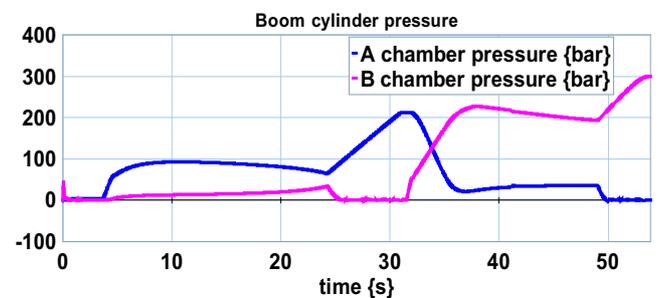


Figure 10: Boom cylinder pressures

The initial pressure at the cylinder is set at 1bar and the maximum limited pressure is 300bar.

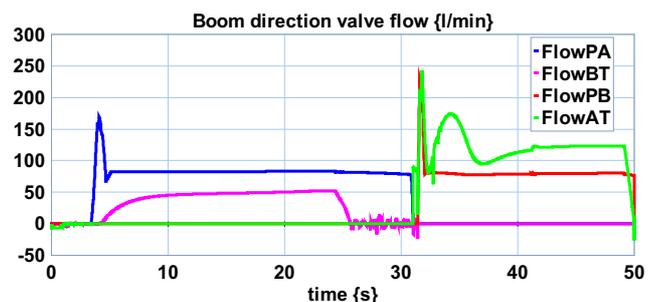


Figure 11: Flow through directional valve

The maximum flow through the proportional direction valve is 200 l/min limited by its opening area. The oscillations of the flow through the directional valve are due to the lack of damping in the cylinder when the end positions are reached.

As mentioned the model can be designed according to the simulation needs. The hydraulic characteristics can be plotted at any point inside the flow circuits. The running results for the slew motor and jib cylinder circuits are not presented due to the size limitation of the paper. It benefits for the crane designers and operators to dimension a crane and to monitor the performance of its hydraulic system.

CONCLUSION

In the previous chapters, a modeling approach using BG method for an offshore crane hydraulic system is described. BG method as a general modeling tool for physical systems allows for modeling of complex systems in multiple energy domains. The 20-sim software provides several modeling libraries and toolboxes for modeling systems of different energy domains. More important, it provides the flexibility to allow the users to build customized models and libraries. In fact, the models in 20-sim hydraulic library are not used because some specific components of the offshore crane system are not included, and the models are hard to understand or edit by the users. The presented hydraulic models of the crane system are built from scratch using basic BG elements.

As part of the future work, a more comprehensive hydraulic library will be developed for offshore machinery systems for modeling and simulation of hydraulic, mechanical, and thermal aspects. A dynamic model of the crane is another part of future work including a 3D animation scenery for visualization. It is also intended to include a hydro-dynamic ship model for study of the impacts of the waves to crane operations.

REFERENCES

- ASSOFLUID, Italian Association of Manufacturing and Trading Companies in Fluid Power Equipment and Components. 2007. *Hydraulics in industrial and mobile applications*.
- Aridhi, E.; Abbes, M.; Mami A. 2013. "Pseudo bond graph model of a thermo-hydraulic system". *2013 5th International Conference on Modeling, Simulation and Applied Optimization* (Apr. 28-30), ICMSAO, Hammamet, 1-5.
- En Qiao Jiang, Shu Han Wang, Jing Ke Du, Xiang Yang Xu, Wen Yong Li . 2013. "Simulation of the Nozzle Electro-Hydraulic Servo Valve with ITI-SimulationX". *Advanced Materials Research*, vol. 690-693, May 2013, 2912-2917.
- Pedersen E.; Engja H.. 2008. *Mathematical Modeling and Simulation of Physical Systems*. Lecture notes in course TMR4257 Modeling, Simulation and Analysis of Dynamic Systems, Department of Marine Technology, Norwegian University of Science and Technology.
- Věchet, S.; Krejsa, J. 2009. "Hydraulic arm modeling via Matlab Simhydraulics". *Engineering Mechanics*, vol. 16, 2009, 287-296.
- Yungang Li; Pengcheng Wang; Liqun Ai; Xiaoming Sang and Jinglong Bu. 2011. "Study on Hydraulic Circuit Simulation Based on MSC·EASY5 for the Arm of Excavator". *Advanced Materials Research*, vol. 291-294, Jul. 2011, 2281-2286.

AUTHOR BIOGRAPHIES

Yingguang Chu studied Mechanical Engineering and Automation in Beijing Technology and Business University, China and obtained his bachelor degree in 2007. After three years working as a production engineer at Sonyericsson, China, he started a master study program in Hydraulic Engineering at Ocean University of China in 2010. After one year he moved to Ålesund, Norway, where he received his master degree in Product and System Design in 2013 and now continues as a research assistant and Ph. D. candidate at Aalesund University College.

Vilmar Æsøy received his Ph.D. in Mechanical Engineering in 1996 from Norwegian University of Science and Technology. From 1997 to 2002 he worked as researcher in Aker Maritime and R&D manager in Rolls-Royce Marine AS. Since 2002 he works as Associate Professor at Aalesund University College.

Øyvind Bunes received his M.Sc. in Marine Technology from the Norwegian University of Science and Technology in 1996. Until 2002 he was employed as a researcher at MARINTEK until he joined ODIM, later Rolls-Royce, in 2002. He now serves as Senior Principal Engineer at Rolls-Royce Marine AS. He is also an Associate Fellow in the Rolls-Royce Engineering Fellowship and an Assistant Professor at Aalesund University College.

Houxiang Zhang received Ph.D. in Mechanical and Electric Engineering in 2003. From 2004 he worked as Postdoctoral Fellow at the Institute of Technical Aspects of Multimodal Systems (TAMS), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Germany. Zhang joined the Department of Maritime Technology and Operation of Aalesund University College, Norway in 2011 and works as Professor on Robotics and Cybernetics.

A real-time UAV INSAR raw signal simulator for HWIL simulation system

Wei Li, Houxiang Zhang, Hans Petter Hildre
Faculty of Maritime Technology and Operations
Aalesund University College
N-6025, Aalesund, Norway
E-mail: {weli & hozh & hh}@hials.no

KEYWORDS

INSAR, HWIL, FPGA

ABSTRACT

In this paper, an FPGA based UAV INSAR raw signal simulator is designed to address high computational complexity. It is based on a time domain raw signal generation algorithm and can compute in real time. This signal simulator is designed for Hardware-in-the-loop (HWIL) UAV INSAR simulation which can be used for UAV operator training and system verification. Multi-FPGAs are used in this simulator with optimisation methods to improve FPGA resource costs, including a modified non-restoring algorithm for slant range computing, as well as pipelined FFT and IFFT processors for a fast convolution method.

INTRODUCTION

Interferometric Synthetic Aperture Radar (INSAR) systems are special types of radar that produce three dimensional high resolution images (comparable to optical sensors) in all weather conditions, night and day (Madsen, S. N et al.1998). High performance INSAR systems utilise sophisticated signal processing algorithms and need complicated and costly radar electronics and processing units.

In recent years, a new remote sensing technology based on Unmanned Aerial Vehicle (UAV) INSAR has emerged. This technology provides a great potential for detailed monitoring and surveillance of areas covering of up to a few thousand square kilometres with a relatively low cost. However, the high performance UAV INSAR is challenging because of the highly dynamic platform and the baseline error caused by antenna oscillation. The raw signal simulator (A. Mori et al. 2004) is useful for testing and verifying the function and performance of the UAV INSAR system.

SAR raw signal generating algorithms can be classified in two types: frequency domain (FD) and time-domain (TD) (G. Franceschetti 1998 and A. Mori et al. 2004). The TD algorithm can easily consider the real trajectory of the platform and other effects such as mechanical structure oscillation and orbital deviations, considerable variation of the velocity vector in the case of an UAV platform which is ideal for a Hardware-in-the-loop (HWIL) simulation system with real time measurement

methods.

However, most TD INSAR simulators do not work in real-time because of the high computational load for the raw signal generating algorithm and the ultra low delay requirement. FPGA (Field-programmable gate array) is an ideal device for implementing the HWIL simulator, thanks to the advantages of high parallel computing performance, low latency and flexible I/O interfaces.

In this paper, multiple FPGAs are used to design a UAV INSAR raw signal simulator for the HWIL simulation system. In order to achieve real-time processing, some optimisation is presented including slant range calculation, fast convolution processor and memory distribution methods. The paper is organised as follows: first, the HWIL simulation system is introduced; second, the INSAR TD raw signal algorithm is described; third, the system architecture of simulator is presented; fourth, the processor design for the raw signal generation is proposed with FPGA design results.

HWIL UAV INSAR simulation system

The UAV INSAR raw signal simulator can be used for HWIL UAV simulation systems, as shown in Figure 1. It is mainly used to design and verify the function and performance of UAV INSAR systems. In this HWIL simulation system, parts of the virtual model are replaced by the actual physical model. For example, the UAV motion simulator is used to simulate UAV dynamics model and is responsible for sending commands and parameters related to the simulation. The INSAR raw signal generator is used to simulate the raw signal according to the radar and platform parameters. The UAV INSAR processor is connected with these devices and works just like in the real environment.

The INSAR target signal simulator is a key part of this HWIL simulation system, which has real-time generation of SAR raw signal and closed-loop UAV simulations. Different from ordinary radar simulation systems, the INSAR raw simulation system needs a large amount computation and a high accuracy. As such, the SAR simulator has strong computing performance and a highly parallel and optimised software design to take full advantage of the hardware computing resources.

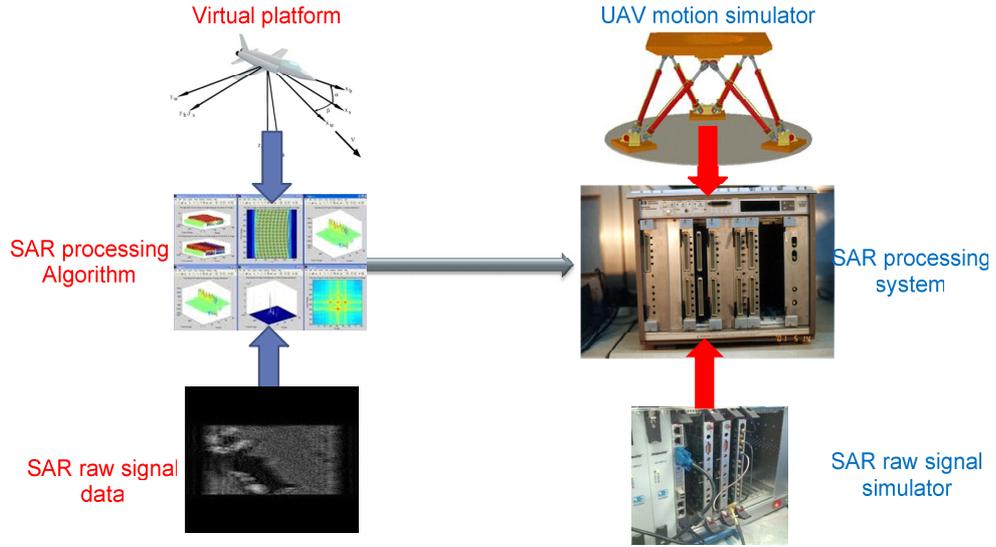


Figure 1: SAR HWIL simulation system

INSAR raw signal simulation algorithm

The geometry of an INSAR system is shown Figure 2, where S1 and S2 are the master and slave antennae. Every PRT, the radar transmits pulse signal in each azimuth position. The raw signal received by the antennae can be described by Eq. 1:

$$s_r = \sum_{i,j} G_{ij} RCS_{ij} \exp[-j4\pi R_s / \lambda] \cdot \delta[r - R_s] \otimes \frac{2}{cv} s_t \quad (1)$$

where

i, j index of sampling in azimuth and range

RCS_{ij} backscattering coefficient of scatter (i, j)

G_{ij} pattern antenna weight

λ wavelength

R_s slant range between a point scatter and the antenna phase centre

$\delta(\cdot)$ pulse envelope

c speed of light

v velocity of platform

s_t transmitted signal

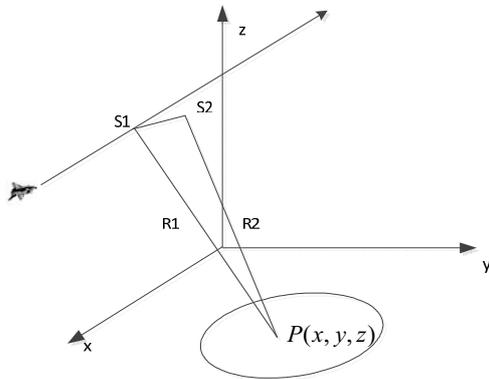


Figure 2: Geometry of an INSAR system

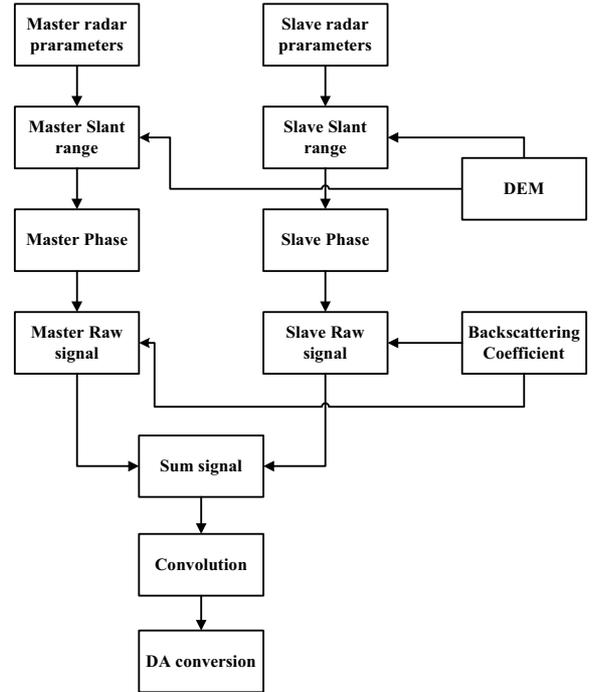


Figure 3: INSAR raw data generating procedure

The raw data generating procedure is based on the block diagram shown in Figure 3. The steps are: 1) Receiving the simulation or test parameters, such as the platform position and velocity; 2) Computing the slant range of scatters in the scene for master and slave antennae; 3) Computing the azimuth phase and multiplied with the RCS; 4) Accumulating the return signal of the same range cell; 5) Convolution with the transmit signal; 6) Converting the digital signal to an analogue signal.

SYSTEM ARCHITECTURE

In order to satisfy the need for processing the TD INSAR raw signal simulation algorithm in real-time, in this paper, multiple FPGAs are used in parallel. The simulator provides master/slave channels and each channel signal is generated using one computing board. As shown in Figure 4, the simulator is based on CompactPCI bus architecture with one main computing board, one slave computing board, an analogue signal generating board and a controller board. The controlling communication is based on PCI bus architecture and the high speed transfer of data between the modules is made possible via backplane LVDS bus architecture.

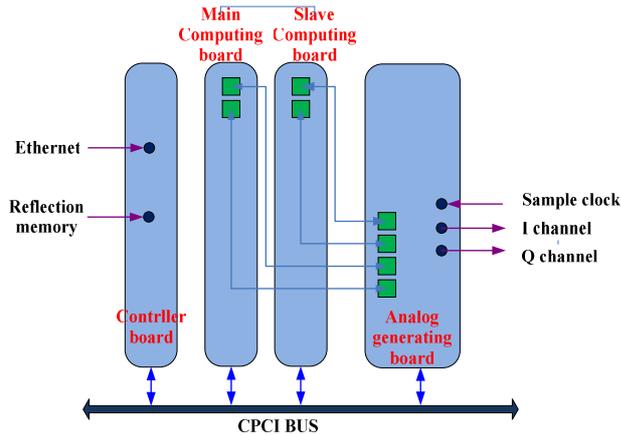


Figure 4: Simulator system architecture

The hardware structure and implementation of the computing module is shown in Figure 5, which was designed and implemented by the authors. The main computing board and slave computing board are of the same structure and are reconfigurable. Five XC5VSX95T FPGAs are used in each computing module and DDRII-SRAM are used as external memory for Digital elevation model (DEM) data and backscattering coefficients because of their high speed, low delay and relatively simplified controller interface. The memory width is 144bits and memory capacity is 16MB for every FPGA. The analogue signal generation module is composed of one XC5VSX95T and one XC4VSX55T which is shown in Figure 6. The FPGA XC4VSX55 is used to transmit simulation parameters to the computing module. The FPGA XC5VSX95T is used to receive the digital raw signal and converts to analogue base band signal through DAC AD9736.

PROCESSOR DESIGN FOR INSAR RAW DATA SIMULATION

According to the system architecture, there is one master and one slave computing module for the digital raw signal generating. The processor architectures are similar in these two modules when multiple FPGAs work in parallel. The main architecture and function of the simulator are shown in Figure 7. The computing boards are mainly used to compute the slant range between radar and targets, phase and sine/cosine value radar return signals. The signals are summed up along

the slant range and then sent to the analogue converting board. The analogue converting board is used mostly for convolution with the transmission signal and conversion to analogue signals using DAC.

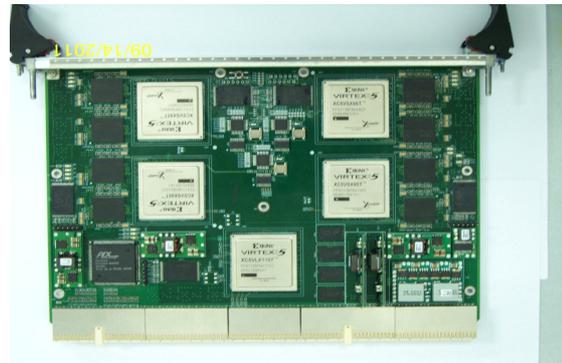
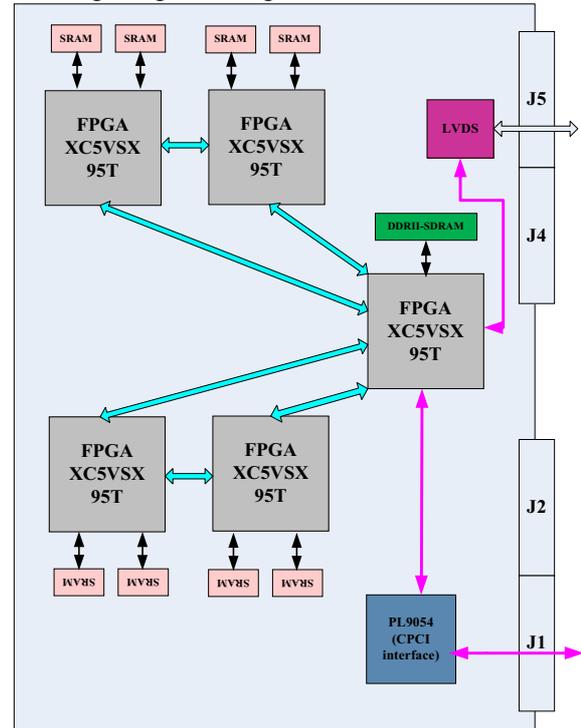


Figure 5: Hardware structure of computing module

In this section, details about the processor design will be described.

Parallel computing architecture

In each computing board, there are four FPGA for main computing and one FPGA for data summing and communication. Each computing FPGA is equipped with four 36bit DDRII-SRAM as external memory for the backscattering coefficient and DEM. The whole SAR imaging area is distributed to every FPGA external memory so that every FPGA can work in parallel without excess communication. According to the radar radiation pattern the area can be divided in the range or azimuth direction. In this simulator, the scene is divided into 4 parts in the range direction corresponding with 4 processing FPGAs. Every FPGA has 24 cores to compute the raw signal. When the processing of signal coherent accumulation is complete, the result of every

processing unit needs to be accumulated for each processing module.

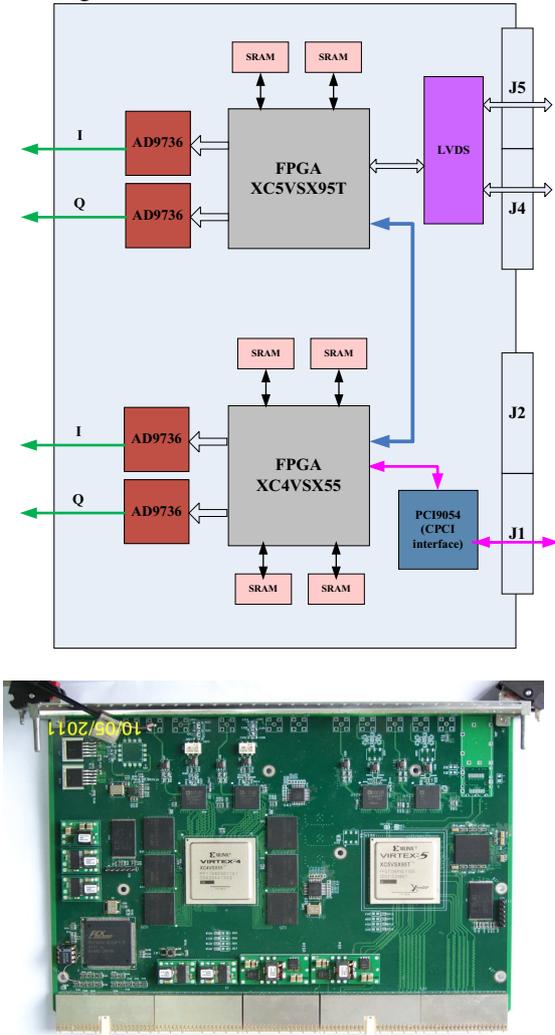


Figure 6: Hardware structure of analogue signal generation module

Slant range computing is one of the main computing steps with highest logic resource cost of FPGA. It is essential for high quality INSAR raw signal generating with high phase accuracy.

$$R_s = \sqrt{(x - X_0)^2 + (y - Y_0)^2 + (z - Z_0)^2}$$

$$\Phi_s = \exp[-j4\pi R_s / \lambda] \quad (2)$$

As shown in Eq. (2), the phase error is directly related to the slant range error. For example, with the centre frequency of 10GHz and the phase error of 3 degrees, the range error should be below 0.0012 metres. If the slant range is computed using (2), a double floating point square root operation is needed. The FPGA resource cost for floating point processing is heavy and the latency greatly increases with precision. As such, the square root is calculated in fixed-point in this paper.

There are mainly three kinds of square root methods, which are Newton-Raphson, SRT-Redundant and non-restoring techniques. In this paper, a modified non-restoring pipelined architecture is used to optimise the hardware resource usage by taking advantage of the FPGA internal CLB structure, which is optimised for adder realisation and by using the RTL approach directly. In every pipelined stage of the non-restoring algorithm the adder and subtractor can be multiplexed using a complement adder. The hardware resource report lists of the 64 bit input square root processors are shown in Table 4 according to the Xilinx ISE12.4 using FPGA XC4VSX55. It can be seen that the LUTs decrease when comparing processors (T.Sutikno.2011 and S. Samavi 2008).

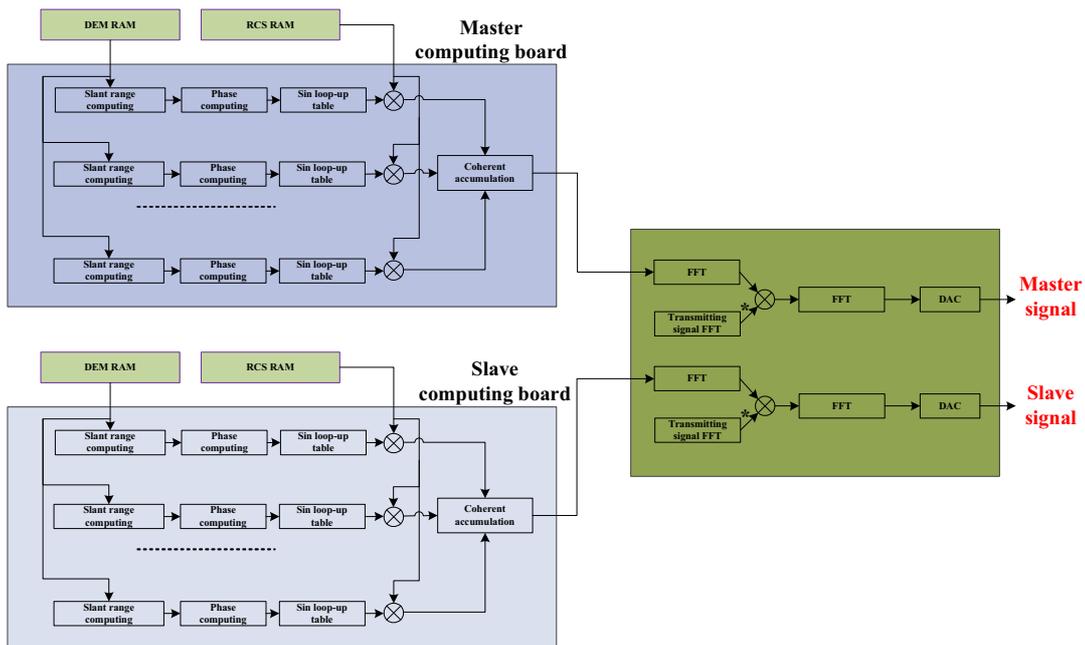


Figure 7: Implementation architecture for raw signal simulation

Table 1 Resource report

Resource Use	Proposed
Slice Flip Flops	1150
Slice LUTs	648
Slices	637

Convolution method

As shown in Figure 8 the signal processed by the main computing modules needs to be convoluted with the transmitted radar signal to generate the raw signal. The convolution can be performed by direct time domain convolution or the fast convolution method using FFT. The time domain convolution method needs lots of resources but has less latency when the operation can be performed in parallel. The resources needed of the latter are reduced, but the latency for the first output is increased. In this paper R2²SDF pipelined FFT processor (S. He et al. 2001) are chosen because of the

high speed and medium resource cost. It can process N FFT points in N clock periods. The R2²SDF output order is bit-reversed. If the same structure is used for the FFT and IFFT, additional memory is required to reorder the output result of the FFT and the latency is greatly increased. So the DIT and DIF structures are presented for the FFT and IFFT processors respectively, which are shown in Figure 9 and Figure 10. The hardware resources – especially memory– decrease when comparing the convolution processor using only DIF FFT. The resource report list is shown in Table 2 according to the Xilinx ISE12.4 using an FPGA XC5VSX95T.

Table 2: Resource report

Resource Use	used
Slices	2340
Block RAM (18Kb)	24
DSP48E	40

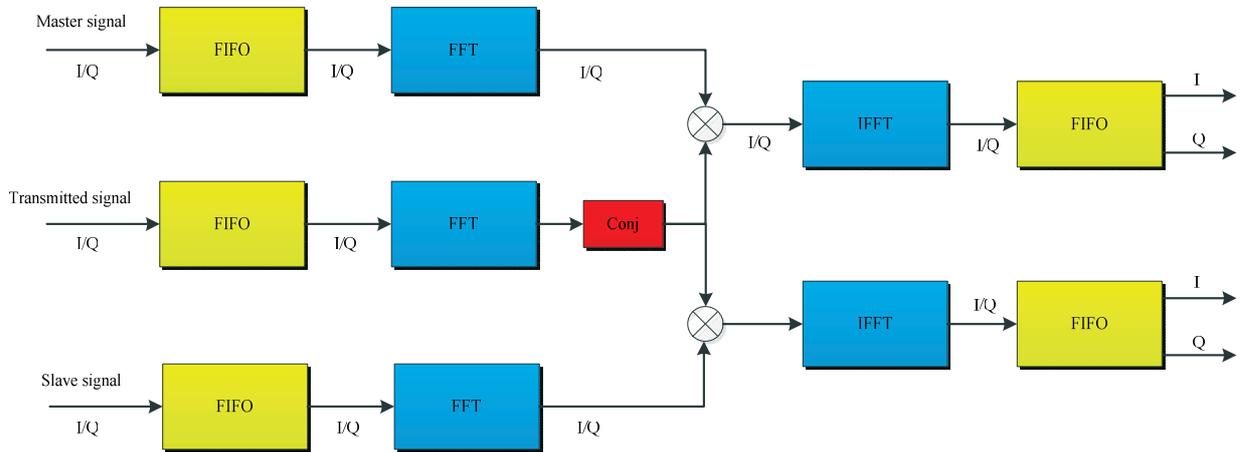


Figure 8: Convolution method

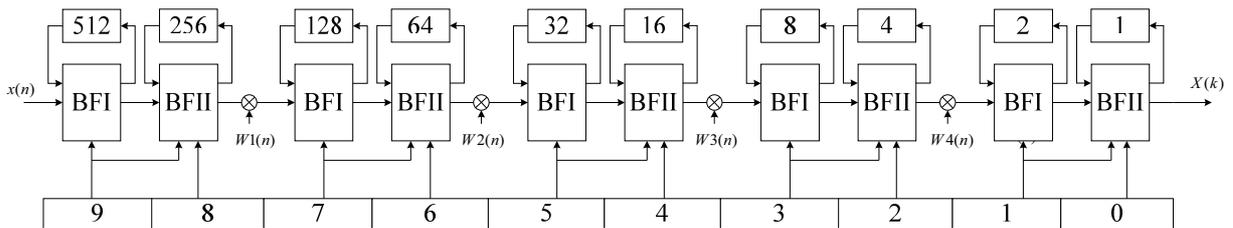


Figure 9: R2²SDF DIF structure

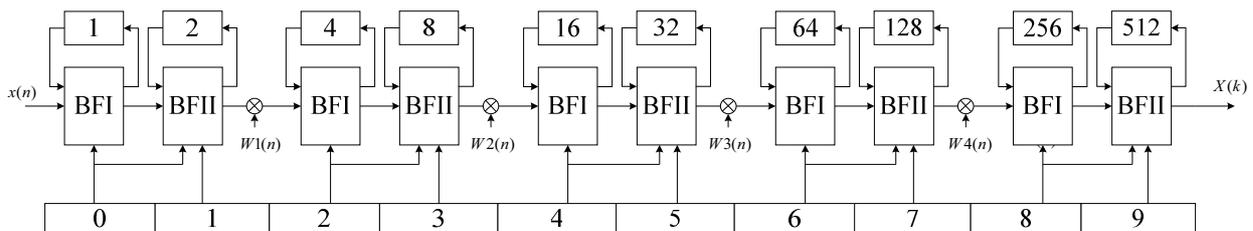


Figure.10: R2²SDF DIT structure

Analogue signal generation

The analogue signal generation is based on FPGA and DAC. The raw SAR analogue signal is generated according to the time radar signal sequence. The PRF pulse signal is a global synchronisation signal for radar transmission and raw signal generation. After the raw signal convolution, the signal data is stored in the FPGA buffer and is generated according to the PRF and slant range centre. In order to generate the analogue signal and compute simultaneously, dual ping-pong rams are used. As shown in Figure 11, when the computing signal data is stored in RAM A, the data from RAM B is read and sent to DAC, while in the following PRF, the order is reversed.

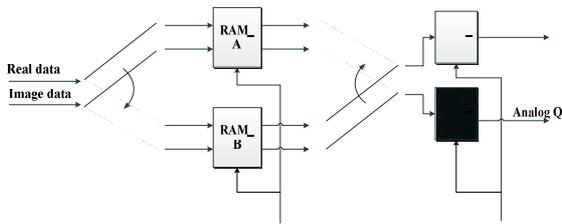


Figure.11: Analogue signal generation structure

FPGA performance

As shown in Figure 6, there are two main kinds of FPGA modules in the system. One is the master/slave computing module the other is the analogue signal generating module. We designed the FPGA modules according to the radar parameters listed in Table 5.

On the computing board, there are four FPGAs for main computing and one FPGA for data summing and communication. Every main computing FPGA has 24 processors which can carry out pipelining tasks, including range computing, phase computing, signal accumulation and memory communication.

We use the Xilinx ISE12.4 for synthesis, place and route. The cost for the main computing FPGA is listed in table iii and the clock speed is 100MHz.

Table 3: Resource report of main computing FPGA

Logic Utilization	Used	Available	Utilization
Registers	35383	58880	60%
Slice LUTs	34358	58880	58%
Block RAM	125	244	49%
Number of DSP48Es	364	640	56%

On the Convolution board, there is one XC5VSX95T for convolution and analogue signal generation, and one XC4VSX55 for communication. The cost for the

convolution and analogue signal generating FPGA is listed in table iii and the clock speed is 100MHz.

Table 4: Resource report for the convolution FPGA

Logic Utilization	Used	Available	Utilization
Registers	19471	58880	33%
Slice LUTs	18279	58880	31%
Block RAM	46	244	18%
DSP48Es	66	640	10%

The simulation test is performed with the parameters listed in Table 5. The continuous computing time of every part is below one PRF (Pulse Repetition Time 25us) and needs less than 51.2ms for 2048-pulse simulation.

Table 5: INSAR simulation parameters

Frequency (GHz)	10
Azimuth points	400
Range points	400
Pulse width (μ S)	1.5
Bandwidth(MHz)	100

CONCLUSION

In this paper we designed an FPGA-based system for real time INSAR raw signal generating. The simulator is based on the TD algorithm and optimisation methods are presented, such as the slant range computing algorithm, parallel processing architecture and the convolution method. It is designed for UAV INSAR HWIL simulation system which can be used for UAV operator training and system verification. In the future the system will be improved by including real-time backscattering coefficient computing and shadow effects.

REFERENCES

- Madsen, S. N., and Zebker, H. A. (1998). Principles and Applications of Imaging Radar. Wiley.
- Ian G. Cumming, Frank H. Wong. (2005). Digital Processing of Synthetic Aperture Radar Data. Artech House.
- Z. Xujin, Z.Zhaoda. (2007). SAR Echo Simulation Based on Hardware-in-loop. Modern Radar. Vol.29, Sept. 2007
- Zheng Xiang, Kaizhi Wang, Xingzhao Liu, Wenxian Yu. (2009). A GPU based Time-domain Raw Signal Simulator for Interferometric SAR. IGARSS 2009, 25-28.
- F Zhang, Zheng Li, B Wang, M Xiang and W Hong. (2012). Hybrid general-purpose computation on GPU (GPGPU) and computer graphics synthetic aperture radar simulation for complex scenes. International Journal of Physical Sciences, 7, 1224-1234.
- Zhang F, Bai L, Hong W. (2008). INSAR imaging geometry simulation based on computer graphics. ISPRS 2008, 781-784.
- Zhang F, Wang BN, Xiang MS. (2010). Accelerating INSAR

- raw data simulation on GPU using CUDA. IGARSS 2010, 2932-2935.
- Zhihua He, Feng He, Zhen Dong, Diannong Liang. (2012). Real-Time Raw-Signal Simulation Algorithm for INSAR Hardware-in-the-Loop Simulation Applications. *IEEE Geosci. Remote Sensing Lett*, 9, 134-138.
- G. Franceschetti, A. Iodice, M. Migliaccio, and D. Riccio. (1998). A novel across-track SAR interferometry simulator. *IEEE Trans. Geosci. Remote Sensing*, 36, 950–962.
- A. Mori and F. De Vita. (2004). A time-domain raw signal simulator for interferometric SAR. *IEEE Trans. On Geosci. Remote Sensing*, 42, 1811–1817.
- L. Yamin and C. Wanming. (1997). Implementation of Single Precision Floating Point Square Root on FPGAs. *IEEE Symposium on FPGA for Custom Computing Machines*, Napa, California, USA, 1997, 226-232.
- T.Sutikno. (2011). An efficient implementation of the non-restoring square root algorithm in gate level. *Internantional Journal of Computer Theory and Engineering*, 3, 1793-8201.
- S. Samavi, et al. (2008). Modular array structure for non-restoring square root circuit. *Journal of Systems Architecture*, 54, 957-966.
- S. He, M.Torkelson. (2001). Designing pipeline FFT processor for OFDM (de)modulation. *ISSSE 2001*, 257-262.
- L.Wei, W. jun, L. shaohong. (2007). The GPS code acquisition based on pipelined FFT processor. *The Second International Conference on Space Information Technology 2007*.

AUTHOR BIOGRAPHIES

WEI LI is a researcher on signal processing at the Faculty of Maritime Technology and Operations, Aalesund University College, Norway. Email: weli@hials.no.

HOUXIANG ZHANG is a professor on Robotics and Cybernetics at the Faculty of Maritime Technology and Operations, Aalesund University College, Norway. Email: hozh@hials.no.

HANS PETTER HILDRE is a professor on product and system design at the Faculty of Maritime Technology and Operations, Aalesund University College, Norway. Email: hh@hials.no.

JIOP: A JAVA INTELLIGENT OPTIMISATION AND MACHINE LEARNING FRAMEWORK

L. I. Hatledal, F. Sanfilippo, H. Zhang

Department of Maritime Technology and Operations
Aalesund University College
Postbox 1517, 6025 Aalesund, Norway

KEYWORDS

Optimisation methods; Machine learning; Object oriented programming; Inverse kinematics

ABSTRACT

This paper presents an open source, object-oriented machine learning framework, formally named *Java Intelligent Optimisation (JIOP)*. While *JIOP* is still in the early stages of development, it already provides a wide variety of general learning algorithms that can be used.

Initially designed as a collection of existing learning methods, *JIOP* aims to emphasise commonalities and dissimilarities of algorithms in order to identify their strengths and weaknesses, providing a simple, coherent and unified view. For this reason, *JIOP* is suitable for pedagogical purposes, such as for introducing bachelor and master degree students to the concepts of intelligent algorithms.

The problems that *JIOP* aims to solve are initially discussed to demonstrate the need for such a framework. Later on, the design architecture and the current functions of the framework are outlined. As a validating case study, a real application where *JIOP* is used to minimise the cost function for solving the inverse kinematics (IK) of a *KUKA* industrial robotic arm with six degrees of freedom (DOF) is also presented. Related simulations are carried out to prove the effectiveness of the proposed framework.

INTRODUCTION

Machine learning refers to the ability of a computer system, or more generally, a machine, to learn from examples (Simon 2013). Essentially, this ability concerns the task of computing a mathematical rule that generalises a relationship initially provided by a finite sample of real data, without being explicitly programmed but by following some kind of training process. This idea is based on the fundamental concepts of representation, evaluation and generalisation. Representation of data instances and evaluation of the same instances by using some kind of assessing function are essential steps for processing the information carried out by the data. Generalisation is a fascinating property that guarantees good system performance on unseen data instances. The set of all possible outputs when given all possible inputs is too large to be covered by the set of observed data. Hence, the system must generalise from the given set of examples, so as to be able to produce useful output in new cases.

Machine learning is highly pervasive today and it is employed in a wide variety of tasks and successful applications in several fields, such as computer vision, natural language processing, pattern recognition, search engines, bio-informatics, robotics, and more generally, for a wide variety of optimisation processes. Optimisation problems often represent very complex tasks and non-heuristic methods are greatly limited in finding proper solutions (Pluhacek et al. 2013). A vast number of different algorithms have already been presented in previous literature and they can be classified into a taxonomy based on the type of input available during the training process (Hormozi et al. 2012). Supervised learning algorithms are trained on labelled examples and their main objective consists of generating a function that maps inputs to desired outputs (Kotsiantis et al. 2007). Unsupervised learning algorithms operate on unlabelled examples and attempt to discover some kind of structure in the data (Alpaydin 2004). Semi-supervised learning combines both labelled and unlabelled examples to generate an appropriate function or classifier (Zhu 2006). Transduction methods try to predict new outputs based on specific and fixed test cases from observed specific training cases (Alpaydin 2004). Reinforcement learning is concerned with learning how to act given an observation of the environment to maximise some notion of reward (Alpaydin 2004). Learning to learn is a model of inductive bias learning based on previous experience (Evgeniou & Pontil 2004).

These machine learning algorithms are often quite complex to implement from scratch and to use properly and efficiently for students, especially considering the limited amount of time that they can spend on this task during their bachelor or master courses of study. As such, it would be very useful to dispose of some kind of developing tool that is easy to use and at the same time gives students a chance to experience the benefit of using machine learning algorithms in practical applications. In particular, a software framework that collects different machine learning methods would greatly help students to emphasise commonalities and dissimilarities of algorithms in order to identify their strengths and weaknesses.

For these reasons, an open source object-oriented machine learning framework, formally named *Java Intelligent Optimisation (JIOP)* has been developed at Aalesund University College to help bachelor and master degree students use existing machine learning algorithms, combine them, extend them or even experiment with new ones. The object-oriented approach was justified by the need to create a clear modular structure for the framework. Moreover, this choice makes it

easy to maintain and modify software. Consequentially, *Java* was chosen as the language of development because it is object-oriented, easy to learn and platform-independent. *JIOP* already provides the following algorithms: Genetic Algorithm (GA) (Deb et al. 2002), Simulated Annealing (SA) (Aarts & Korst 1988), Differential Evolution (DE) (Storn & Price 1997), Particle Swarm Optimisation (PSO) (Kennedy 2010) and Artificial Bee Colony (ABS) (Karaboga & Basturk 2007). However, new methods can be easily developed and added to the framework. The framework is available under a Berkeley Software Distribution (BSD) license and can be retrieved from the following website: <https://github.com/aauc-mechlab/JIOP>.

RELATED RESEARCH WORK

In recent years, the machine learning community has developed a notable number of different libraries. However, most of the time, these libraries are specifically designed for a particular algorithm and for a defined application. Moreover, they are typically written by using different languages and only a few of them are publicly available. Very few comprehensive collections of different algorithms are freely available to developers and students.

In (Kohavi et al. 1994), Kohavi et al. introduced *MLC++*, a library of C++ classes for supervised machine learning. *MLC++* (up to version 1.3.X) is in the public domain and is still distributed as such by the *Silicon Graphics International* (SGI) Corporation. SGI *MLC++* (V2.0 and higher) includes improvements to the original *MLC++*. However, even if these improvements are available in both source and object code formats, they are only in the research domain. In (Abeel et al. 2009), Abeel et al. presented *Java-ML*, a collection of machine learning and data mining algorithms for both software developers and research scientists. The interfaces for each type of algorithm are quite clear and algorithms strictly follow their respective interface. In (Heaton & Reasearch 2010), Heaton outlined an advanced machine learning framework, formally named *Encog*. This framework supports a variety of advanced algorithms, as well as support classes to normalise and process data. Most *Encog* training algorithms are multi-threaded and scale well to multi-core hardware. *Encog*, which is available for Java, .Net and C/C++, can also make use of a Graphical Processing Unit (GPU) to improve processing time.

It should be noted that these libraries are often created to be used by professional developers and as such, do not have a very strong pedagogical orientation.

SYSTEM ARCHITECTURE

The *JIOP* design architecture aims to provide simplicity and flexibility. The general idea is that the user should be able to use the already-implemented algorithms or even implement new algorithms with the least amount of effort. In order to accomplish this, the framework relies on a number of abstract classes and well defined interfaces to do most of the work. Moreover, the framework supports generic types, denoted by an $\langle E \rangle$ in this paper, which allows the user to choose how to represent the variables to optimise. Thanks to generics, the user is not only in control of whether or not the variables should be stored in an array, a list, or some other user defined type, but also if the variables should be stored as doubles, floats,

strings, etc. Also, *JIOP* provides multi-threading support on a selected set of functions, more specifically functions that creates and evaluates multiple candidates in a single method call. A Unified Modelling Language (UML) class diagram showing the software architecture is available in Fig. 1. The general idea is that all algorithms must extend the base class, the *MLAlgorithm*, whose goal is to optimise its candidates' variables. A candidate is an *Object* with a set of variables stored in an encoding instance and a cost which relates to the fitness of the variables. A *CandidateFactory* instance may be used to create new candidates. The following subsections gives a more detailed description of the *JIOP* classes.

Base classes

The abstract *MLAlgorithm* class, described in Table I, is the base class of the framework and defines a single abstract method that subclasses must implement in order to function as a *JIOP* optimisation class. The purpose of the abstract *internalIteration()* function in all subclasses is to optimise *Candidate* instances in a single step. Furthermore, this class keeps a reference to the best *Candidate* found and also keeps a history of the performance using the *MLHistory* class.

The abstract *PopulationBasedMLAlgorithm* class is a subclass of *MLAlgorithm* and adds functionality to store a population of *Candidate* instances, as well as keeping a *MLHistory* of the average performance.

The *Candidate* class, described in Table II, is a wrapper around an *Encoding* instance and a corresponding cost. The cost is a measure of performance, where a low cost is desirable. Moreover, as the *Candidate* class implements the *Comparable* interface from the Java standard library, any array or *Collection* of *Candidates* can be sorted based upon its affiliated cost. A related subclass is the final *EvaluatedCandidate* class, which has additional information on time used and the number of iterations that were necessary to find the candidate solution.

The *Encoding* interface is the basic template for any class used to hold variables. The methods that need to be implemented are shown in Table III. More specifically, classes that implement this interface must store the actual variables used in the optimisation process. The user may choose any Java *Object* to represent the variables due to generics and should implement this interface accordingly. However, some implementations are included in the framework. Currently, these are implementations for *double[]*, *float[]*, *List<Double>* and *List<Float>* encodings. This general interface is extended by another interface, *ParticleEncoding*, which is a special case used for PSO optimisation, as this optimisation technique has a velocity associated with its variables (Kennedy 2010).

The abstract *EncodingFactory* class defines a set of factory methods, all of which returns an *Encoding* instance. These methods must be implemented by subclasses, and are subsequently used by the *CandidateFactory* class, described in Table V, to create new *Candidate* instances.

The *CandidateContainer* class extends the *ArrayList* class from the Java standard library and is used to store candidates. In addition to the standard list functionality, this class provides functions for sorting, printing and getting the average score of the contained candidates.

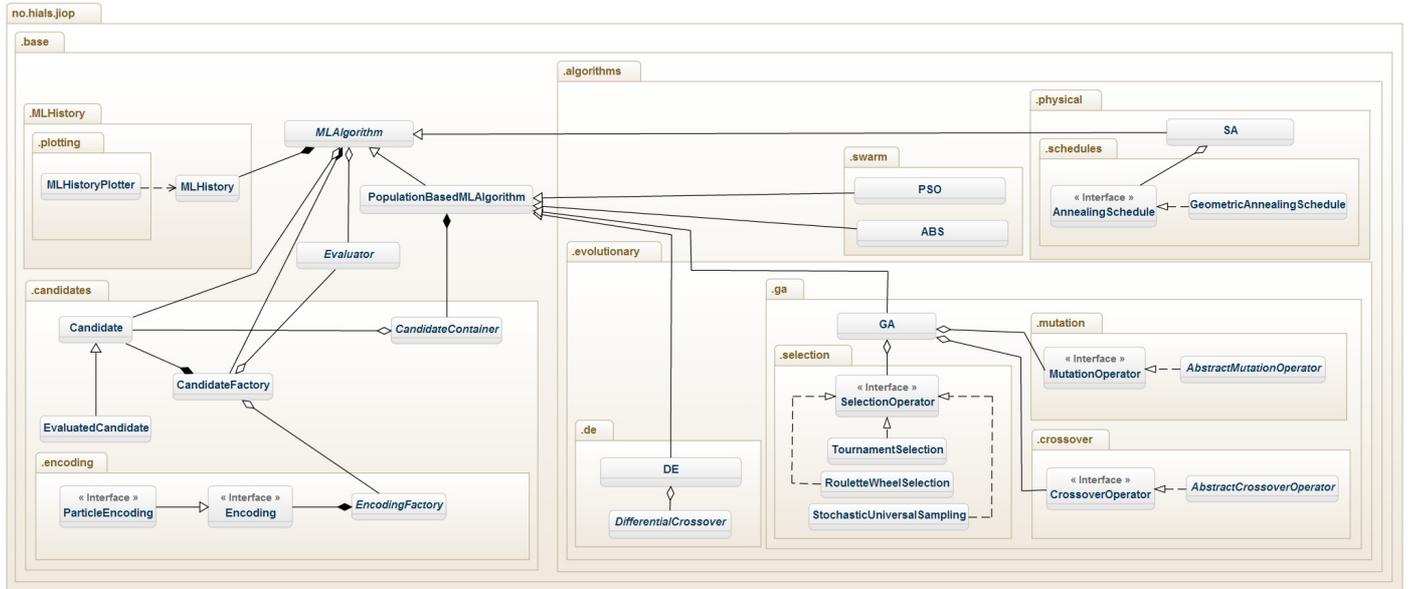


Fig. 1. A UML class diagram depicting the JIOP core classes in their respective software package

TABLE I. THE MLALGORITHM CLASS

Abstract class <i>MAlgorithm</i> <E>	
Data members -Candidate<E> bestCandidate -AbstractEvaluator<E> evaluator -CandidateFactory factory -MLHistory history -ExecutorService pool Functions +EvaluatedCandidate<E> runfor(long millis) +EvaluatedCandidate<E> runFor(double error) +EvaluatedCandidate<E> runFor(int iterations) +void iteration() +void reset(boolean clearHistory) +void reset(boolean clearHistory, List<E> seed) +void writeHistoryToFile(String location) +void submit(List<Runnable> jobs) <i>Setters and getters omitted...</i> Abstract functions +void internalIteration()	A reference to the best found candidate A reference to the evaluator A reference to the candidate factory A reference to the optimisation history Thread pool used for multi-threading Runs the algorithm for the specified amount of time Runs the algorithm until the error falls below the specified error Runs the algorithm for the specified number of iterations Invokes <i>internalIteration()</i> and updates the <i>MLHistory</i> Resets the algorithm and populates it with random candidates Resets the algorithm and populates it with new candidates based on the seed. Writes the history to a text-file Uses multi-threading to finish the submitted jobs Does a single optimisation step, implemented in a subclass

TABLE II. THE CANDIDATE CLASS

Class <i>Candidate</i> <E> implements Comparable	
Data members -BasicEncoding<E> encoding -double cost Functions +Encoding<E> getEncoding() +E getVariables() +double getCost(double cost) +void setCost() +int compareTo(Candidate c)	The candidate encoding. The candidate cost Returns the encoding Returns the encodings variables Returns the cost of the candidate Updates the cost Used to sort the candidates

TABLE III. THE ENCODING INTERFACE

Interface <i>Encoding</i> <E>	
E getVariables() int size() Encoding<E> copy()	The variables - double[], String[] etc. The number of variables A unique copy of the instance

TABLE IV. THE EVALUATOR CLASS

Abstract class <i>Evaluator</i> <E>	
Functions +double evaluate(Candidate<E> candidate) +double evaluate(Encoding<E> encoding) Abstract Functions +double evaluate(E variables)	Evaluates the candidate Evaluates the encoding Evaluates the variables

historic perspective of a *MAlgorithm*. The class constructor takes the iteration number, time-stamp and the cost as a input, and adds the data to a list, which can be used later for plotting.

The *MLHistoryPlotter* uses a 3rd party library, *JMathPlot* (μ -labs, 2012), and derives from the *Plot2DPanel* from said library. This class populates a line graph depicting the method's performance with regards to time or iterations, based on the *MLHistory* data from a supplied *MAlgorithm*.

Optimisation classes

The *DE* class is an implementation of the Differential Evolution algorithm. It optimises a problem by maintaining a population of *Candidate* instances and creates new ones by

The abstract *Evaluator* class, described in Table IV, defines a single abstract method that subclasses must implement in order for the evaluation process to work. Furthermore, it has two extra convenience methods for evaluation.

The *MLHistory* class is a basic class in charge of storing a

TABLE V. THE CANDIDATEFACTORY CLASS

Class <i>CandidateFactory</i> <E>	
Data members -ExecutorService pool -Evaluator<E> evaluator -EncodingFactory<E> factory Functions +Candidate<E> random() +Candidate<E> toCandidate(E e) +Candidate<E> neighbour(Candidate<E> c) +List<Candidate<E>> randomCandidates(int n) +List<Candidate<E>> toCandidates(List<E> e) +List<Candidate<E> neighbourCandidates(Candidate<E> original, int n)	Thread pool used for multi-threading A reference to the evaluator A reference to the encoding factory Returns a candidate with random variables Returns a candidate based on the variables Returns a neighbour candidate Returns a List of n candidates (multi-threaded operation) Creates a list of candidates from the list of variables (multi-threaded operation) Returns a list of n neighbour candidates (multi-threaded operation)

combining existing ones. The behaviour of the algorithm can be influenced by modifying the size of the population NP , the weighting factor F and the crossover weight CR .

The *PSO* class is an implementation of the Particle Swarm optimisation algorithm using a swarm of *Candidate* instances. This method implements the *ParticleEncoding* interface as it allows the *Candidate* to be updated according to the PSO scheme. The behaviour of the algorithm can be influenced by modifying the size of the swarm, the inertia weight ω and the learning factors c_1 and c_2 .

The *ABS* class is an implementation of the Artificial Bee Colony algorithm, which is based on the intelligent behaviour of a honey bee swarm. The behaviour of the algorithm can be influenced by setting the size of the colony and the number of bees employed as scouts.

The *GA* class is an implementation of a continuous Genetic algorithm. It holds a population of *Candidate* instances and optimises a problem by mimicking natural evolution using *elitism*, *selection*, *crossover* and *mutation*. These operators can be implemented by the user by way of the *SelectionOperator*, *CrossoverOperator* and *MutationOperator* respectively. The behaviour of the algorithm can be greatly influenced depending on the selection and mutation operators as well as the population size, crossover rate, mutation rate and elitism variables.

The *SA* class is an implementation of the Simulated Annealing algorithm. It tries to mimic the annealing process in metallurgy. It does not rely on a population of *Candidate* instances, but uses a current *Candidate* and generates neighbours of this instance. Furthermore, it has a starting temperature and uses an annealing schedule to regulate it, defined by the *AnnealingSchedule* interface, and a standard acceptance probability function.

CASE STUDY

The case study focuses on finding a solution to the IK problem, which consists of determining the joint parameters that provide the desired position and orientation of the *KUKA* robot's end-effector. A solution to this problem can be found using classical approaches, such as analytically using the Jacobian matrix or through geometrics. However, the geometric approach does not scale well with the number of DOFs to be controlled, as the complexity of the calculations increases. Furthermore the Jacobian approach is known to have stability issues around singular configurations due to matrix inversions. Therefore, modifications to the classical Jacobian must be introduced. The main advantage of using machine learning

algorithms is to save the user from having to hard-code geometric equations or deriving the Jacobian matrix from the model's forward kinematics (FK). Moreover, singular configurations are not ill-posed as no matrix inversions are performed. In order to conduct this case study, the same framework that our own research group introduced in (Sanfilippo et al. 2013), is used. More specifically, *JIOP* is used within this framework.

Description of the evaluation function

The evaluator measures the cost of the candidates, and is made up of three components:

- 1) Positional error.
- 2) Orientation error.
- 3) The change in joint angles between two consecutive IK solutions.

The sum of these components gives the cost of a proposed candidate solution, where a lower score is better.

In this case study, the variables in the *Candidate* instances are stored as a *double[]*, which is an array of double precision floating point numbers. A single value represents a joint angle θ_n , whereas the whole array represents a complete joint configuration $[\theta_1, \theta_2, \dots]$ for the *KUKA* robot. FK is utilised to compute the resulting end-effector position and orientation given a set of angles.

The position cost is found using the euclidean norm:

$$a = \sqrt{(p_{1x} - p_{2x})^2 + (p_{1y} - p_{2y})^2 + (p_{1z} - p_{2z})^2} \quad (1)$$

where p_1 and p_2 are the normalised desired and candidate position vectors.

The orientation cost is also found using the euclidean norm:

$$b = \sqrt{(o_{1x} - o_{2x})^2 + (o_{1y} - o_{2y})^2 + (o_{1z} - o_{2z})^2} \quad (2)$$

where o_1 and o_2 are the normalised desired and candidate orientation vectors.

The cost related to the changes in joint angles between two successive solutions is given by (3) and is the sum of the per-element absolute difference between the previous solution $\hat{\theta}(t-1)$ and the candidate solution $\hat{\theta}(t)$.

$$c = |\hat{\theta}(t-1) - \hat{\theta}(t)| \quad (3)$$

This cost is added to encourage similar solutions, because the 6-DOF *KUKA* robot is redundant and can have multiple valid

TABLE VI. ALGORITHM SPECIFIC PARAMETERS

Genetic Algorithm		Particle Swarm Optimisation		Differential Evolution		Artificial Bee Colony		Simulated Annealing	
Population Size	100	Swarm size	40	Population size	30	Swarm Size	30	Starting temperature, t_0	100
Selection size	.5	Inertia weight, ω	.9	Weighting factor, F	.8	Scouts	6	Annealing Schedule	Geometric
Mutation rate	.5	Local bias, c_1	.9	Crossover rate, CR	.9				
Elitism	.1	Global bias, c_2	.9						
Selection	SUS								

solutions.

The cost returned by the cost function is then given by:

$$cost = \alpha a + \beta b + \gamma c \quad (4)$$

where α , β and γ are weighting factors. In this case study, these are: $\alpha = 1$, $\beta = 1$ and $\gamma = 0.03$, where α and β were chosen to consider the position and orientation error equally, while γ was chosen by trial and error.

Algorithm parameters

The algorithm-specific parameters used in this case study are shown in Table VI. For the GA algorithm, the crossover rate is the probability of recombination, the mutation rate is the probability of a mutation occurring at the gene level, the elitism is the percentage of the population that survives unaltered into the next generation and the population size is the number of individuals to use. The selection type used by the GA is an implementation of *Stochastic Universal Sampling* (SUS).

The parameters used in the PSO algorithm are the inertia weight, the learning factors and the swarm size. The inertia weight controls the velocity, the learning factors are biases toward the local and global best positions respectively and the swarm size is the number of particles to use.

For the DE algorithm, the weighting factor controls the amplification of differential variation, the crossover weight probabilistically controls the amount of recombination while the population size is the number of parameter vectors to use.

The ABS algorithm parameters are the number of bees in the colony, and the number of bees employed as scouts. The scouts are responsible of looking for promising food sources and communicate findings to the rest of the swarm.

Finally, the starting temperature for the SA algorithm is the initial temperature of the system. The SA uses a *GeometricAnnealingSchedule* with a constant decay rate of 0.9 to iteratively cool the temperature.

Optimisation using the JIOP framework

After extending the *Evaluator* class and implementing the *evaluate(E variables)* function, an instance of this class is passed on to the constructor of an *MLAlgorithm* along with a *CandidateFactory* and additional algorithm-specific methods and parameters. If the algorithm is population based, then a *CandidateContainer* instance is also required. After instantiating a *MLAlgorithm* instance, the user may call one of the *runFor()* methods, which returns an *EvaluatedCandidate* with the result. For a graphical representation of the result, the user can initiate an *MLHistoryPlotter* and pass the respective algorithm's *MLHistory* instance to it. The resulting plot can then be shown in a graphical window. Alternatively, the user may write the data to a text-file, using the *writeHistoryToFile()*, for plotting in some external software.

SIMULATION RESULTS

In this section, the performance of the machine learning algorithms currently found in the *JIOP* framework are presented. In particular, these are a DE, a PSO, a ABS, a GA and a SA implementation. In order to simulate a real operational scenario of the *KUKA* robot using position control, a set of points that defines a possible trajectory for the end-effector was chosen. In this operational scenario, couples of adjacent points do not differ much from each other statistically. Moreover, the algorithms have no more than a 50 ms time frame to produce a solution in order to maintain a real-time control scenario. Starting from the beginning of the second solution, the initial population of the respective algorithms is injected with a fraction of the previously best found candidate solutions, which is a feature of the *MLAlgorithm* and mimics elitism in GAs. Fig. 3, 4, 5, 6 and 7 shows a plot of the cost versus time of the different algorithms. The data used to produce the plots is gathered from the respective algorithms' *MLHistory* and saved as a text file, using the *dumpDataToFile()* function, and then plotted using MATLAB. Table VIII shows the resulting position and orientation of the end-effector. It is clear that the machine learning algorithms implemented in *JIOP* and presented in this paper are able to find the solution to the given problem quickly and accurately. It should be noted that a cost of zero is not possible because of the third component in the cost function, given by (4). Furthermore, Fig. 2 shows the resulting poses of the calculations.

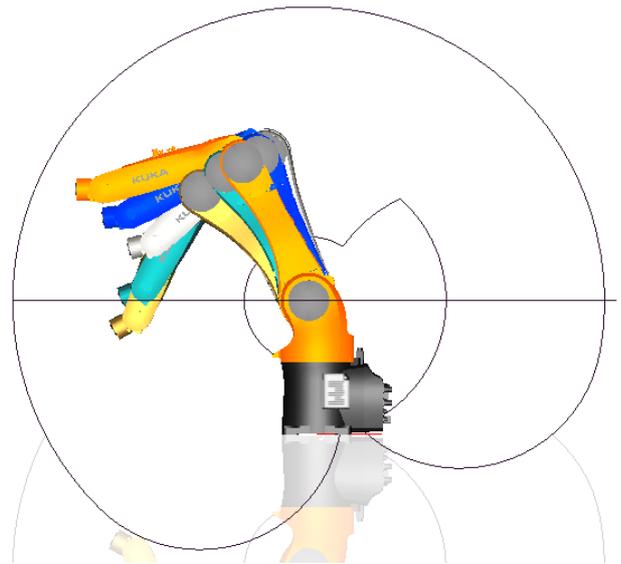


Fig. 2. Visualisation of the five resulting *KUKA* poses, found using the *JIOP* framework.

The computation was done on a computer running Windows 7 with a quad core Intel Core i7-3820QM processor. The result of utilising the processor's multiple threads is given in Table VII, and shows the total number of iterations that the different algorithms were able to complete in the given time-frame. It is clear that utilising multiple threads is highly beneficial to the performance of the algorithms. Note that the Simulated Annealing implementation runs on a single thread. The result for four threads is therefore undefined.

TABLE VII. MULTI-THREADING PERFORMANCE

Algorithms	Iterations	
	1 thread	4 threads
Differential Evolution	1508	4229
Particle Swarm Optimisation	1095	3375
Artificial Bee Colony	871	2181
Genetic Algorithm	581	1852
Simulated Annealing	42305	-

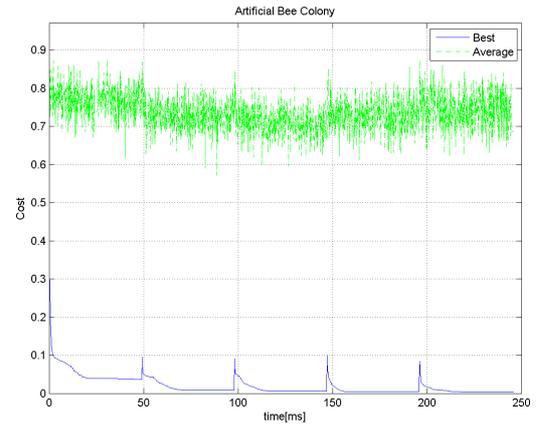


Fig. 5. Cost versus time using the Artificial Bee Colony algorithm from the *JIOP* framework solving five consecutive IK solutions

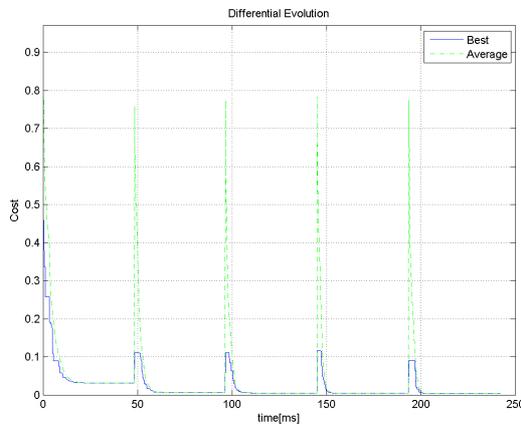


Fig. 3. Cost versus time using the Differential Evolution algorithm from the *JIOP* framework solving five consecutive IK solutions

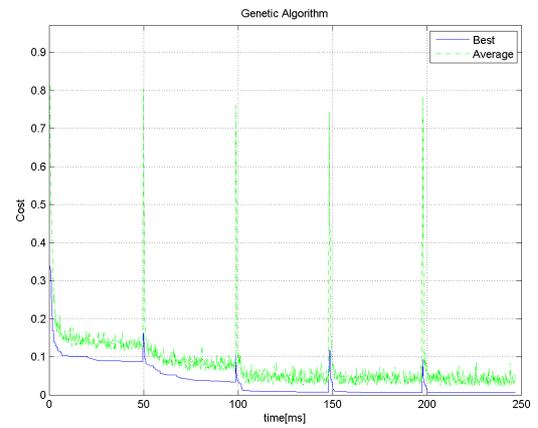


Fig. 6. Cost versus time using the Genetic Algorithm from the *JIOP* framework solving five consecutive IK solutions

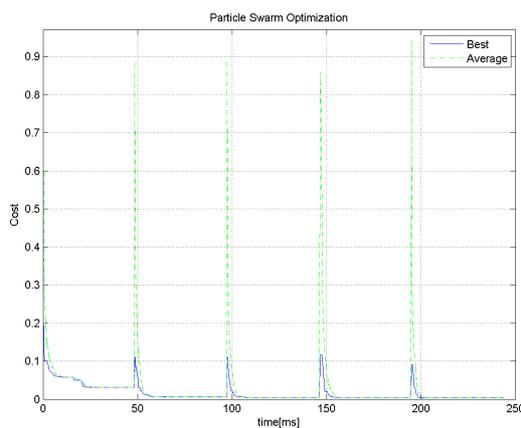


Fig. 4. Cost versus time using the Particle Swarm Optimisation algorithm from the *JIOP* framework solving five consecutive IK solutions

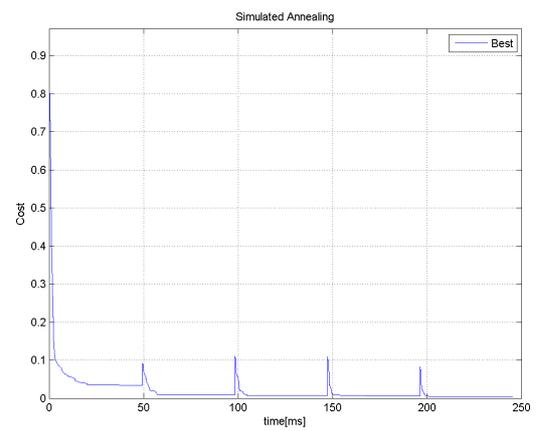


Fig. 7. Cost versus time using the Simulated Annealing from the *JIOP* framework solving five consecutive IK solutions

TABLE VIII. SIMULATION RESULTS

	Position [m]	Orientation [deg]	
Desired	[0.800, 0.000, 0.600]	[0.000, 0.000, 0.000]	
	[0.700, 0.000, 0.500]	[0.000, 7.500, 0.000]	
	[0.600, 0.000, 0.400]	[0.000, 15.000, 0.000]	
	[0.600, 0.000, 0.250]	[0.000, 22.500, 0.000]	
	[0.600, 0.000, 0.150]	[0.000, 30.000, 0.000]	
DE	[0.800, -0.000, 0.600]	[0.000, 0.000, -0.000]	0.0314
	[0.700, 0.000, 0.500]	[-0.000, 7.500, -0.000]	0.0066
	[0.600, -0.000, 0.400]	[-0.000, 15.000, -0.000]	0.0052
	[0.600, 0.000, 0.250]	[-0.000, 22.500, -0.000]	0.0044
	[0.600, -0.000, 0.150]	[-0.000, 30.000, -0.000]	0.0030
PSO	[0.800, -0.000, 0.600]	[0.000, -0.000, 0.000]	0.0314
	[0.700, 0.000, 0.500]	[0.000, 7.500, -0.000]	0.0066
	[0.600, 0.000, 0.400]	[-0.000, 15.000, -0.000]	0.0052
	[0.600, -0.000, 0.250]	[0.000, 22.500, -0.000]	0.0044
	[0.600, 0.000, 0.150]	[0.000, 30.000, 0.000]	0.0030
ABS	[0.801, -0.000, 0.601]	[-0.130, -0.136, -1.236]	0.0370
	[0.700, 0.000, 0.500]	[-0.048, 7.489, -0.076]	0.0083
	[0.600, 0.000, 0.400]	[0.001, 15.027, -0.021]	0.0055
	[0.600, 0.000, 0.250]	[-0.002, 22.513, -0.014]	0.0046
	[0.600, 0.000, 0.150]	[0.015, 29.987, -0.004]	0.0031
GA	[0.881, 0.019, 0.667]	[-0.031, 0.031, -0.025]	0.0887
	[0.702, 0.006, 0.501]	[-0.066, 7.591, -0.357]	0.0332
	[0.600, 0.000, 0.400]	[-0.093, 14.950, -0.273]	0.0071
	[0.600, 0.002, 0.252]	[-0.011, 22.544, -0.055]	0.0061
	[0.599, 0.002, 0.155]	[-0.008, 30.002, -0.007]	0.0054
SA	[0.801, 0.003, 0.602]	[-0.178, 0.006, -0.010]	0.0340
	[0.700, 0.001, 0.501]	[-0.419, 7.522, -0.097]	0.0086
	[0.601, 0.001, 0.401]	[0.238, 15.006, -0.081]	0.0070
	[0.601, -0.001, 0.248]	[0.131, 22.577, 0.054]	0.0064
	[0.600, 0.001, 0.151]	[-0.143, 29.871, 0.114]	0.0046

CONCLUSION AND FUTURE WORK

JIOP demonstrates an object-oriented machine learning framework to anyone with an interest in machine learning algorithms and the Java programming language. This is especially true for students that are in a need of a compact, easy to use and highly configurable framework that also provides visual feedback. In this case, *JIOP* delivers an effective and lightweight environment for using and creating machine learning algorithms. A continuous effort will be made to streamline and expand the framework with even more algorithms and configuration options.

REFERENCES

Aarts, E. & Korst, J. (1988), ‘Simulated annealing and boltzmann machines’.

Abeel, T., Van de Peer, Y. & Saeys, Y. (2009), ‘Java-ml: A machine learning library’, *The Journal of Machine Learning Research* **10**, 931–934.

Alpaydin, E. (2004), *Introduction to machine learning*, MIT press.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. (2002), ‘A fast and elitist multiobjective genetic algorithm: Nsga-ii’, *IEEE Transactions on Evolutionary Computation* **6**(2), 182–197.

Evgeniou, T. & Pontil, M. (2004), Regularized multi-task learning, in ‘Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 109–117.

Heaton, J. & Reasearch, H. (2010), ‘Encog java and dotnet neural network framework’, *Heaton Research, Inc.*, Retrieved on July **20**, 2010.

Hormozi, H., Hormozi, E. & Nohooji, H. R. (2012), ‘The classification of the applicable machine learning methods in robot manipulators’, *International Journal of Machine Learning and Computing* **2**, 560–563.

Karaboga, D. & Basturk, B. (2007), ‘A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm’, *Journal of global optimization* **39**(3), 459–471.

Kennedy, J. (2010), Particle swarm optimization, in ‘Encyclopedia of Machine Learning’, Springer, pp. 760–766.

Kohavi, R., John, G., Long, R., Manley, D. & Pflieger, K. (1994), Mlc++: A machine learning library in c++, in ‘Proceedings of the Sixth International Conference on Tools with Artificial Intelligence’, IEEE, pp. 740–743.

Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. (2007), ‘Supervised machine learning: A review of classification techniques’.

Pluhacek, M., Senkerik, R., Zelinka, I. & Davendra, D. (2013), Multiple choice strategy for pso algorithm - performance analysis on shifted test functions., in ‘ECMS’, European Council for Modeling and Simulation, pp. 393–397.

Sanfilippo, F., Hatledal, L. I., Schaathun, H. G., Pettersen, K. Y. & Zhang, H. (2013), A universal control architecture for maritime cranes and robots using genetic algorithms as a possible mapping approach, in ‘Proceeding of the IEEE International Conference on Robotics and Biomimetics (RO-BIO) Shenzhen, China, December 2013’, IEEE, pp. 322–327.

Simon, P. (2013), *Too Big to Ignore: The Business Case for Big Data*, Wiley. com.

Storn, R. & Price, K. (1997), ‘Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces’, *Journal of global optimization* **11**(4), 341–359.

Zhu, X. (2006), ‘Semi-supervised learning literature survey’, *Computer Science, University of Wisconsin-Madison* **2**, 3.

AUTHOR BIOGRAPHIES

LARS IVAR HATLEDAL received a bachelor’s degree in Automation from Aalesund University College, Norway. Here, Hatledal joined the Department of Maritime Technology and Operations as a Project Leader in June 2013. Email: laht@hials.no

FILIPPO SANFILIPPO is a PhD candidate in Engineering Cybernetics at the Norwegian University of Science and Technology, and a research assistant at the Department of Maritime Technology and Operations, Aalesund University College, Norway. He obtained his Master’s Degree in Computer Engineering at the University of Siena, Italy. Email: fisa@hials.no

HOUXIANG ZHANG received Ph.D. degree in Mechanical and Electronic Engineering in 2003. From 2004, he worked as Postdoctoral Fellow at the Institute of Technical Aspects of Multimodal Systems (TAMS), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Germany. Dr. Zhang joined the Department of Maritime Technology and Operations, Aalesund University College, Norway in April 2011, where he is a Professor in Robotics and Cybernetics. Email: hozh@hials.no

ENHANCING UNDERGRADUATE RESEARCH AND LEARNING METHODS ON REAL-TIME PROCESSES BY COOPERATING WITH MARITIME INDUSTRIES

Webjørn Rekdalsbakken¹

Filippo Sanfilippo²

¹ Department of Information and Communication Technology

² Department of Maritime Technology and Operations

Aalesund University College

N-6025 Aalesund, Norway

E-mail: wr@hials.no

KEYWORDS

Collaborative Learning, Undergraduate research, Real-time process control, Image processing, Embedded systems.

ABSTRACT

Building embedded real-time systems of guaranteed quality, in a cost-effective manner, raises challenging scientific and technological problems. For several years at Aalesund University College (AAUC), there has been ongoing activity in the development of embedded real time systems in close cooperation with private technology developers from the local industry. Much of this work is related to the design and development of systems for process control and camera surveillance of industrial processes, with an emphasis placed on operations on board ships. The main purpose is to maintain and improve safety and efficiency in industrial and ship operations. A very effective way to meet this goal consists of developing distributed embedded real time systems that independently monitor different dedicated tasks and are integrated in a common ubiquitous network. In this context, bachelor students at AAUC are involved in several research projects that give them the possibility of working in an industrial prospective scenario and under the supervision of both their professors and company-employed engineers. Most often, these activities are also part of an innovative educational and research loop, in which the projects evaluated as having the greatest innovative potential are followed up with new prototypes and research activities. By adopting a pedagogical prospective, this paper introduces an overview of the most promising student projects and methods for including bachelor students in real-time process control research activities is presented. The pedagogical and technological bases of this work are first discussed, and afterwards an analysis is done regarding the importance of COTS (Commercial Off-The-Shelf) products and system integration. Software and communication protocols are discussed from a system integration point of view. Finally, the results and conclusions of this approach are presented.

INTRODUCTION

Aalesund University College (AAUC) is located in the region of Møre og Romsdal, Norway. A particular challenge for this region, which is interested in facilitating the growth of technology-based industries and business start-ups, is the conceptualisation and promotion of the linkages between the regional science base, reflected most heavily in the region's universities but also in its non-academic research laboratories, and its industry base. AAUC has been working with remote autonomous embedded systems for process control and camera inspection purposes on board ships (Rekdalsbakken and Osen, 2012) over the past several years. Some of these projects have played a critical part in inspiring new concepts for more efficient and secure work operations on board ships and as such have been implemented in broader research programs in cooperation with the industry. Many of these projects have been realised as embedded systems on different hardware platforms. In this activity, modern sensors together with vision technology and wireless communication systems are combined into self-contained systems independently performing important tasks in ship operations. Most of these systems originated in student projects from the "Real-Time Programming" course or from other courses in the Automation Engineering Program (Sanfilippo et al. in press) and in the students' bachelor theses. A systematic evaluation and selection was done on these projects to pick out the most promising ideas and products with potential for further study. Follow-up work has been done on some concepts in the form of further development, done in steps of gradual improvement over time until their results could either be continued as industry projects or academic research programs. Therefore, some of these activities have initiated the development of industrial products or the publication of scientific papers.

PEDAGOGICAL BASIS

It is important to build enthusiasm from the start when working in this kind of educational environment. The objectives are to foster motivation and feelings of self-

reliance, to build team spirit, and to help the students develop the skills to seek and retrieve the relevant information. The team learning approach, with students working in groups, constitutes the basis in the process of defining, adapting and selecting appropriate projects. The process involves both the university staff and external partners. The students are encouraged to present their own ideas and proposals in order to include the newest technologies and program tools in their projects. In this way the students are allowed to work with realistic problem set-ups by using the most recent advances in technology and methods. The students are involved from start to finish in this immersive approach. The idea originally comes from Piaget and focuses on operational learning. It is outlined in the book *Piaget in school* (Hundeide, 1985). The basis of this view lies in the concept of *implicit learning*, which represents a concept of knowledge based on experience. In this view, emphasis is placed on the significance of practical experience and close contact with real-life and applied research. In such a view, the student laboratory, as a working place, will have a central and challenging position. Piaget's theories about learning emphasises that the knowledge must be made personal through acting. By active processing of real situations the knowledge acquires meaning. Piaget calls the result of this process *operational knowledge* in contrast to figurative knowledge, which is the representation by the senses of the external situation (Kleive et al., 1994). This strategy can also be viewed as a kind of *Action Science* or *multimodal teaching and learning* (Kress et al., 2001) or as a process of building new knowledge through acting and the reflection about acting. This method is extremely demanding for the teacher, who must act more as a stimulator than as an authority and has to arrange for optimal conditions of active learning. According to our experience at AAUC on project based learning, the role of the teacher should be more that of a catalyst and a counsellor. In such a situation it is important that the students be confident that the teachers have the necessary professional skills to define the correct problem settings and to give appropriate advice and corrections concerning the students' work.

TECHNOLOGICAL BASIS

In this setting, it is important that the university staff work in a suitable context for selecting the most promising project ideas for the students. To be of benefit for the students, the projects must be relevant to the field of study of the students, have a realistic scope and be possible to complete in the given time. There are several steps during the development and study process, including: orientation, research framework and expectations discussion, weekly and monthly meetings, mid-term presentations and assessments. Furthermore, the technology and methods to be used must be among the most recent in the market. The fields of consumer electronics and game technology are the most relevant fields in this context, and they also represent the most rapidly changing current technologies. This is a

demanding situation, but it also has the potential to come up with new ideas and to develop new and flexible products. This means that the student projects usually include an extended use of novel and advanced components, regarding electronics, micro-controllers, sensors, cameras and data acquisition equipment. Since problem settings are chosen from an application point of view, most projects are also related to the local maritime industry. Speaking from a theoretical point of view, it is possible to generalise by saying that the projects belong to one of three categories. The first is *stabilisation and control of motion platforms*, the second is *object recognition and surveillance* by use of cameras, and the third is *remote and autonomous control strategies* for all sorts of robotic vehicles. These three categories are analysed in the following sections.

Motion platforms

Motion stabilisation represents a challenging problem in all sorts of ship activity. Given that the ship is always in motion when at sea, care must be taken in all kinds of operations to avoid dangerous situations and damage of materials. This is relevant for all kinds of lifting and handling operations, especially when cargo is to be taken through the sea surface. Equipment such as cameras, search lights and precision instruments must also be screened from the ship movements. It is possible to compensate for this unwanted motion by mounting the equipment on a stabilised basis or platform, which counterbalances these movements. The construction of suitable stabilised platforms represents the best method for improving and avoiding these problems. To this end, the theory of kinematics and inverse kinematics is employed. The most common platforms are physical systems of three or six independent axes or degrees of freedom (DOFs). At AAUC, the work on motion platforms mostly concerns the development of a full-scale physical high-speed craft simulator, and stabilisation of equipment on board ships, such as cameras and search lights (Nogva et al., 2008), (Rekdalsbakken, 2005, 2006, 2007). A laboratory model of a 6-DOFs motion platform is shown in Figure 1. Based on a Cartesian coordinate system, the deck can be translated along three independent axes, in addition to being independently rotated about each of these axes. For example, the transformation of a coordinate basis in 3-DOFs from a given position to an arbitrary new position, is given by the transformation matrix below:

$$P_{pr} = \begin{bmatrix} \frac{L}{2} \cos \theta - \frac{\sqrt{3}L}{6} \sin \phi \sin \theta & -\frac{L}{2} \cos \theta - \frac{\sqrt{3}L}{6} \sin \phi \sin \theta & -\frac{\sqrt{3}L}{6} \sin \phi \sin \theta \\ \frac{\sqrt{3}L}{6} \cos \phi & \frac{\sqrt{3}L}{6} \cos \phi & -\frac{\sqrt{3}L}{3} \cos \phi \\ \frac{L}{2} \sin \theta + \frac{\sqrt{3}L}{6} \sin \phi \cos \theta & -\frac{L}{2} \sin \theta + \frac{\sqrt{3}L}{6} \sin \phi \cos \theta & -\frac{\sqrt{3}L}{3} \sin \phi \cos \theta \end{bmatrix}$$

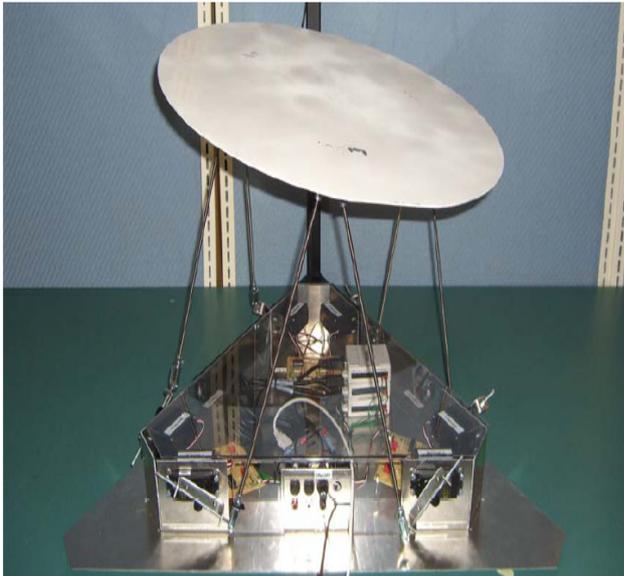


Figure 1: A laboratory Model of a 6-DOFs Motion Platform

Such matrix transformations naturally lead to the search for a general description of the relocation of object coordinates in a coordinate basis for any dimension. This is achieved by describing the objects and their transformations as tensors. The tensor algebra defines general linear transformations in a way that is valid in all kinds of coordinate systems. This is because a tensor includes the transformations of both the object coordinates and the coordinate basis itself. This kind of mathematics is also very useful in the description of the kinematics of 3D graphical objects. To display a 3D object from all directions one will need an extra dimension, i.e. a 4-dimensional coordinate system. The most general approach to obtain this is to define a number system in the equivalent number of dimensions. The obvious benefit is that each point of an object is now defined by a number, and the numbers follow the arithmetic rules of the number system. A 4-dimensional number system was defined by the Irish mathematician and physicist William Rowan Hamilton in 1843 (Graves, 1882). It consists of one real and three imaginary axes. He called it “Quaternions”. Today this number system is widely used by scientists and engineers to describe the general motion of 3D objects.

Surveillance systems

In many situations of potentially dangerous or advanced operations there is a need for close and continuous inspection of the working environment. In many situations, modern camera technologies can be used to extend the natural sight of an operator so that a better survey of a difficult or hardly accessible area can be obtained, for instance in rescue operations. With the aim of making such devices available at affordable prices in a competitive situation, the search for standard open source software and cheap hardware components is crucial to the future realisation of the products. The

experience is also that the frontier of this technological area is driven by developers of open source public domain software and the multitude of producers of small scale hardware systems. The challenge is actually to foresee the technology market and to integrate components and software into useful systems. The intention is to build general camera surveillance systems. Such systems usually consist of two independent parts that communicate over a network as a socket connection with a client-server protocol. The communication may be based on a wireless or a wired LAN connection. The server side of the system is built upon an embedded hardware platform of which the central equipment is a camera that can be controlled to acquire images from any direction in its sight of view (yaw $\pm 180^\circ$, pitch $\pm 90^\circ$). The system is equipped with the necessary devices for fast image analysis in real time and driver software for the network communication. The client system is usually a computer with a network connection and software powerful enough for the communication and presentation of real-time image streams. On this topic, a number of experiments have been performed at AAUC, with different kinds of software and components. Some of these systems are for real applications in offshore operations (Xu, 2011) or dynamical positioning of ships (DP) (Brandal et al., 2011). Other applications are for inspection on land, e.g. the search by mobile robots for pollution, dangerous materials and mines (Fjørtoft and Lund, 2010), (Håheim et al., 2010) or automatic product control (Gao et al., 2009). As these systems become more autonomous and self-sufficient there will be a growing need to combine them into more comprehensive networks in order to compare and group vital information from many sources. This will give a much better overview of complex situations.

Vehicle control

The third main area of interest at AAUC is the development of remotely controlled and autonomous vehicles. Such mobile robots are used for numerous purposes and have a broad and increasingly important position in many kinds of operations. Very often such vehicles combine several advanced technologies into an integrated system. For instance, a robot intended for object search and recognition must have a sophisticated image acquisition and analysis system, inertial sensors for the registration of angles and accelerations and a wireless radio or WiFi system. In addition, a robust control system is needed. In recent literature, the most famous of such vehicles is the Mars rover, but the majority of the other related robots have much less sophisticated duties to attend, e.g. the search for mines in post-war locations. At AAUC, several control systems for such vehicles have been designed using different strategies. To speed up the prototyping process and save time on the mechanical design, model assembly kits with steering and speed control already implemented have been purchased from the consumer market. These kits have been used as bases to build up

new control systems for the robots. The communication usually takes place over a radio communication link or a Wi-Fi network, typically controlled by a cell phone or a game controller. The most widely use of such vehicles is in search and surveillance operations. An example of one of these student projects is an iPhone controlled RC vehicle by use of the Apple remote control program OSCremote (Giske et al., 2009). Moreover, in the last few years, our research group at AAUC, together with our international partners, has also put some effort on proposing new prototypes of mobile robots. For instance, the mobility of several newly designed modular robots has been investigated. In particular, the configuration of a five-limbed modular robot was studied (Liu et al., 2012). A specialised locomotion gait was designed to allow for omnidirectional mobility as shown in Figure 2. Due to the large diversity resulting from various gait sequences, a criterion for selecting the best gaits based on their stability characteristics was proposed. In particular, considering the static stability of the pentapedal configuration different states s can be identified and a set of n gaits can be considered. When the walking direction x is given, a 120×10 matrix $M(x)_{n,s}$ can be derived, where $M(x)_{n,s}$ indicates the stability margin of the s^{th} state of the robot within the No. n considered gait. Moreover, two indices can be used as references for the gait stability. The first is the minimum stability margin $M(x)_{nmin}$, which evaluates the most vulnerable situation to disturbances during the whole period. The second one is the summed stability margin $M(x)_{ntotal}$, which is used to investigate the overall stability gait. A series of simulations was performed to test and evaluate the various gaits in different walking directions. In this way the pentapedal robot was trained to walk in a stable manner. In Figure 3 the simulations results for all the stable gaits in the 90° walking direction are shown.

THE IMPORTANCE OF COTS AND SYSTEM INTEGRATION

One of the most important advantages of this academia-industry cooperation is that these projects represent a way of testing the possible future realisation of new concepts and prototypes, both with regard to technology and economy. Having many groups of students working with similar problem settings on different hardware and software platforms provides an insight into solutions that may be worth taking further in a research project. For the implementation into a real industrial system, e.g. on a ship, among all of its necessary and useful devices, there has to be a motivated demand for that product. Secondly, there must be documentation to prove that the technology is reliable and necessary components will be easily available in the future. Furthermore, software solutions must be openly available and easy to improve and maintain. These kinds of systems must easily fit into the existing solutions for the operation of the ship. To manage the realisation of such systems it is highly important to have a broad knowledge of the electronic

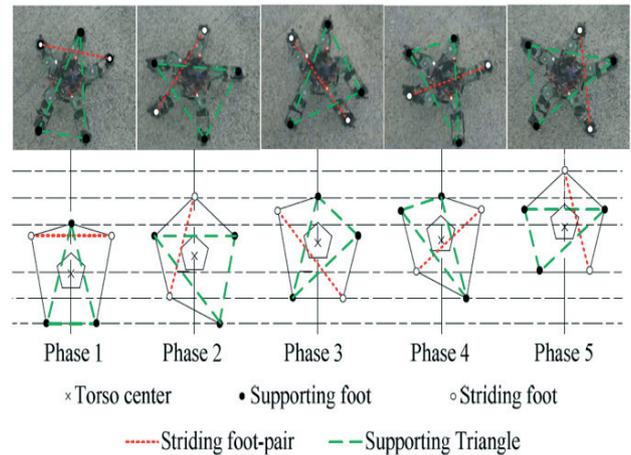


Figure 2: Stable Gait of the Pentapedal Robot

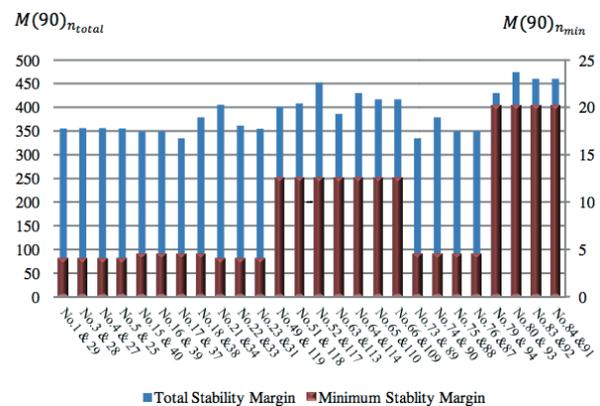


Figure 3: Total and Minimum Stability Margin of all the stable gaits in the 90° walking direction

COTS market, especially in consumer and game products. Because of the large market and multitude of products the frontier of new electronic systems lies here and this market segment represents a vital dynamic. Wireless communication, camera systems and modern sensor technology have had and still have an enormous development through the mobile phone market and game industry. In particular, gyro and accelerometer components of high quality and small size, and cameras of extreme image quality and usability, have been developed in the last few years and can be bought at very low costs. Together with freely available software tools and driver libraries, this situation opens up a new world of exciting challenges to the engineer for developing complex embedded systems for all kinds of innovative applications.

Equipment and Assembly

In all our experiments, the involved computer hardware and all necessary equipment are available at ordinary customer retailers and net shops. A typical example is a student project set-up for building a search vehicle. The search operation is realised by building a small scale

autonomous model vehicle equipped with a web camera which is placed on top of a stabilised motion platform, typically 3-DOFs. The motion platform and the steering system are usually controlled by using an Arduino Uno micro-controller board, featuring an ATmega328 micro-controller from Atmel. Because of the substantial computational load, the image capturing process is implemented by using a Hardkernel Odroid U2 open development platform that sends the image stream over a Wi-Fi connection to a stationary PC. The PC receives the stream and performs the image analysis. When the position of the object is found, its location relative to the vehicle is calculated, i.e. the distance and angle of direction. In particular, the target object is located by utilising two still images or frames taken from different points of view by the robot in such a way that a stereoscopic image of the target can be obtained. The location of the object is sent to the micro-controller on the vehicle and used to calculate the reference signals to control the vehicle and the camera. The internal vehicle application tasks are designed to do all local control and data acquisition, such as stabilising and positioning of the camera, controlling the vehicle motion, and capturing and transmitting the camera images. Everything is built and developed by use of low cost and openly available hardware and software. The system architecture implementation and class diagrams are shown in Figure 4 and Figure 5.

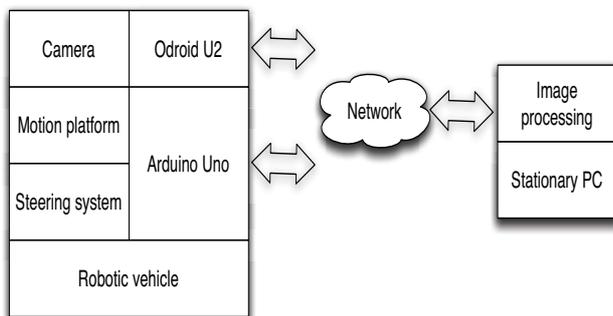


Figure 4: System Architecture

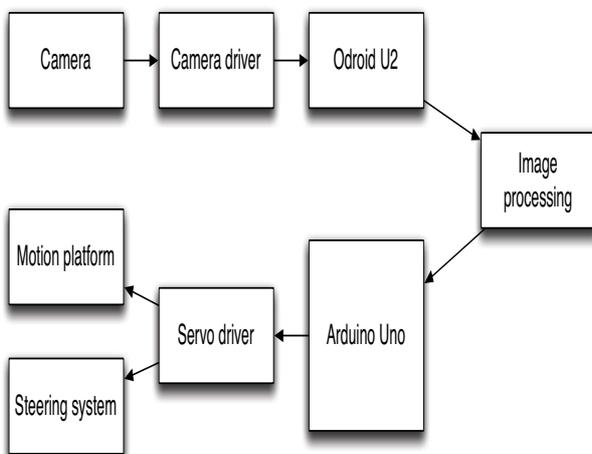


Figure 5: Vehicle Connection Diagram

SOFTWARE AND COMMUNICATION

The software implementation of these kinds of systems is quite extensive and represents the most intensive work and expensive part of the project. Thus it is important to choose a software base and development tools that are easy to work with, easy to update and low-cost. These are good arguments to look for public domain software. In general, Java is the programming language chosen for the software development of these projects, and NetBeans is the preferred development environment. Standard Java is an open source system, free for anyone to download and use. However, for each micro-controller and a chosen operating system, an application programming interface (API) is usually provided. This includes a runtime system and a software development kit. These products often implement a reduced version of standard Java and with some additional libraries for special purposes. In addition to standard Java, some custom tailored program tools and auxiliary program libraries must be used. Most important in hard real-time applications is that the Java runtime engine connects closely to the underlying hardware, which usually is the case with dedicated controllers.

Concurrent Java

Java is the preferred software development system mainly because of its platform independence and extensive and easily implemented real time qualities. It provides an integrated thread mechanism and a broad range of thread management methods. It also has an extended set of timer functions and an effective software interrupt system, implemented as a set of event listeners. The event listeners capture events sent to them from hardware devices via the Java runtime engine.

Linux OS and driver software

It is important to choose a lightweight and flexible operating system. Linux OS exists in many variations from independent developers, and all of them are openly available and free to use. When designing embedded systems, finding suitable driver software for needed equipment is crucial. Developing driver software is a time-consuming task. Linux has great potential in terms of driver software, and the desired software can usually be found on the Internet in public repositories.

Software Libraries

In spite of the extensive application programming interface of Java there will always be a need for special software methods. When working with systems comprising special and extensive equipment, specialised software is needed. Examples of such software applications are acquisition and analysis of images from cameras, and handling of communication protocols for the networks. Software for controlling different kind of motors and actuators will also be necessary. Software methods governing such dedicated areas are often

collected into complete software libraries. Many of these libraries can be found for free on the Internet. A broad insight into the existence of such facilities is very useful and can save many hours of work. Considering the numerous software packages available on the free software market, the ability to choose and adapt the most useful system programs is vital.

RESULTS

In order to assess the effectiveness of the proposed pedagogical and research approach, the students were periodically surveyed and questioned. The students were asked to evaluate the effectiveness of the course structure, project and course elements. The students also had the opportunity to comment on aspects of the course that they valued and aspects of the course that they felt should be revised. A common response from the students at the end of a course is that the project was very important to them and that they wanted more time devoted to it. The results of this pedagogical and research system with continuous selection and improvement of student projects over a period from 2003 to present are in accordance with the goals of the plan, both with regard to students and teachers, as well as for local industry partners. Several project ideas have been tested in cooperation with local industry partners, and relations with local companies have been strengthened. A total of approximately 150 projects have been involved during the past ten years. About 15 scientific publications have been presented on international conferences and some of these have also been published in highly reputable journals. Most projects have been carried out in cooperation with local industry, for the bachelor thesis this amounts to about 80%.

DISCUSSION

This paper presents a systematic way to include bachelor students in engineering programs into research activities. The students belong to the study program in Cybernetics, and the focus is on the course Real-time programming and on the students' bachelor thesis. The primary issue of this work is to investigate new trends in sensor and vision systems for possible integration of such technologies in new products and operations, especially on board future ships. As the tasks required of modern products become increasingly more complex and demanding, there is a growing need for system integration from many fields and insight into the consumer and COTS market. Here, the focus was to explore the rich market of consumer electronics driven mainly by the cell phone and game industry, using these components in the development of advanced and low-cost customer specified products. Experiments with selected equipment of this type were performed mainly by building small scale physical models. However, the aim is to integrate these new technologies into new products for the future. The models only represent an easy approach for testing equipment and methods in an

adequate way for real operations. The tests will reveal and document which of these devices have a potential for further development in an industrial context. By this scheme, the activities best suited to pursue for further projects may be selected. This is like an interactive trial and error process where components and methods are systematically selected for further refinement and development. Through this process all students have a way to gain an insight into product development and scientific methods.

CONCLUSION

The underlying aim of this work is to investigate some new products in the consumer technology market for possible future employment of such technologies in the development of new products, mainly for operations on board ships. This activity of practical testing of vision and sensor systems on small scale models has shown to represent an effective way to reveal the potential for adapting such technologies in modern ship operations. The surplus of new devices within wireless communications, vision systems and MEMS (Micro-Electro-Mechanical-Sensors) components available in the consumer market, driven by the cell phone and game industry, represents fantastic new possibilities in the integration and adaptation of advanced technology in more traditional fields, both in the production process and in the products. The experiments performed in this work have revealed that it is of great benefit for young engineers and applied scientists to have a thorough insight into this market. The results of the experiments also encourage further work towards implementation and use of such technologies in product development. The way of conducting these projects, with careful selection of new ideas, and systematic follow-up of selected projects in close cooperation with industry companies, has also shown to be very fruitful. The industry partners acquire a tool to select and test new product ideas; the students have a chance to make good contacts with local industry companies and training in scientific thinking. Finally, the teachers are able to keep themselves technologically updated and have access to a system for applied research activities.

ACKNOWLEDGEMENT

We would like to express our gratitude to the hard working students at AAUC that participate in these projects.

REFERENCES

- Brandal B, Solibakke T, Tennø E, Øien O. 2011. "Dynamisk posisjonering med Web-kamera". B.Sc. thesis, Aalesund University College, Aalesund.
- Fjortoft, V, Lund M. 2010. "Kamerakontroll for mobil søkerobot". B.Sc. thesis, Aalesund University College, Aalesund.
- Gao X, Sørland T, Xu Q. 2009. "Vision system for automatic quality control". B.Sc. thesis, Aalesund University College, Aalesund.

- Giske R, Vågsæter P K, Årø D. 2009. "iPhone kontrollert RC bil". Prosjekt i Sanntids Datateknikk. Aalesund University College, Aalesund.
- Graves A P. 1882. "Life of Sir William Rowan Hamilton, Volume I". Dublin University Press.
- Hundeide K. 1985. "Piaget i skolen". J. W. Cappelen's Forlag AS.
- Håheim Ø, Saure T, Siqveland M. 2010. "Nettverkstyrt modellbil med stabilisert kamera". B.Sc. thesis, Aalesund University College, Aalesund.
- Kjerstad N, Rekdalsbakken, W. 2006. "Low Cost Three Degrees of Freedom Motion Platform for High-Speed Craft Simulator." Proceedings of Marsim 2006.
- Kleive P. E, Rekdalsbakken W, Årskog V. 1994. "Laboratoriets plass i utdanningen ved HIÅ". Report at Aalesund University College, Aalesund.
- Kress G, Jewitt C, Ogborn J, Tsatsarelis C. 2001. "Multimodal teaching and Learning". British Library Cataloguing-in-Publication Data. ISBN 0-8264-4859-3.
- Liu C., Sanfilippo F., Zhang H., Hildre H. P., Liu C. and Bi S. 2012. "Locomotion Analysis of a Modular Pentapedal Walking Robot." *Proceedings of ECMS 2012, the 26th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9564944-4-3, CD ISBN: 978-0-9564944-5-0.
- Nogva J, Remøyholm R, Stern C. 2008. "Hurtigbåtsimulator". B.Sc. thesis, Aalesund University College, Aalesund.
- Osen, O L., Kristiansen H T., Rekdalsbakken, W. 2010. "Application of Low-Cost Commercial Off-The-Shelf (COTS) Products in the Development of Human-Robot Interactions." *Proceedings of ECMS 2010, the 24th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9564944-0-5, CD ISBN: 978-0-9564944-1-2.
- Rekdalsbakken, W. 2005. "Design and Application of a Motion Platform in Three Degrees of Freedom." In Proceedings of SIMS 2005, the 46th Conference on Simulation and Modelling, pp 269-279, Tapir Academic Press, NO-7005 TRONDHEIM.
- Rekdalsbakken, W. 2006. "Design and Application of a Motion Platform for a High-Speed Craft Simulator." In Proceedings of ICM 2006, IEEE 3rd International Conference on Mechatronics, pp. 38-43, IEEE Catalog Number of Printed Proceedings: 06EX1432.
- Rekdalsbakken, W. 2007. "The Use of Artificial Intelligence in Controlling a 6DOF Motion Platform." *Proceedings of ECMS 2007, the 21th European Conference on Modelling and Simulation*, pp 249-254. European Council for Modelling and Simulation, printed ISBN: 978-0-9553018-2-7, CD ISBN: 978-0-9553018-3-4.
- Rekdalsbakken, W, Styve, A. 2007. "Real-Time Process Control with Concurrent Java." *Proceedings of the 6th EUROSIM Congress on Modelling and Simulation*, pp 120, ISBN-13: 978-3-901608-32-2, ISBN-10: 3-901608-32-X.
- Rekdalsbakken, W. 2007. "Feedback Control of an Inverted Pendulum by Use of Artificial Intelligence." *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 11, No. 9, pp. 75-80. Fuji Technology Press.
- Rekdalsbakken, W. 2008. "Intelligent Control of an Inverted Pendulum." *Intelligent Engineering Systems and Computational Cybernetics*. Springer Verlag. ISBN: 978-1-4020-8677-9.
- Rekdalsbakken, W, Styve A. 2008. "Simulation of intelligent ship autopilots." *Proceedings of ECMS 2008, the 22th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9553018-5-8, CD ISBN: 978-0-9553018-5-5.
- Rekdalsbakken, W, Osen O L. 2009. "Teaching embedded control using concurrent Java." *Proceedings of ECMS 2009, the 23th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9553018-8-9, CD ISBN: 978-0-9553018-9-6.
- Rekdalsbakken, W, Osen O L. 2012. "Exploring artificial vision for use in demanding ship operations." *Proceedings of ECMS 2012, the 26th European Conference on Modelling and Simulation*. European Council for Modelling and Simulation, printed ISBN: 978-0-9564944-4-3, CD ISBN: 978-0-9564944-5-0.
- Sanfilippo, F., Osen O L., Alaliyat S. (in press) "Recycling a discarded robotic arm for automation engineering education." *Proceedings of ECMS 2014, the 28th European Conference on Modelling and Simulation*.
- Xu, Q. 2011. "Real Time Object Recognition System for Offshore operations". MSc. Thesis, Aalesund University College, Aalesund.



AUTHOR BIOGRAPHIES

WEBJØRN REKDALSBAKKEN is Assoc. Professor and leader of the Bachelorprogram in Cybernetics at Aalesund University College (AAUC), Norway. He is also elected Prorector at AAUC. He was the last Rector at the former Møre og Romsdal Engineering College, which was one of three colleges that merged into AAUC in 1994.



FILIPPO SANFILIPPO is a PhD candidate in Engineering Cybernetics at the Norwegian University of Science and Technology (NTNU), and a research assistant at the Department of

Maritime Technology and Operations, Aalesund University College, Norway. He obtained his Master's Degree in Computer Engineering at University of Siena, Italy.

BALLAST WATER ANALYSIS AND HEAT TREATMENT USING WASTE HEAT RECOVERY SYSTEMS ON BOARD SHIPS

Yanran Cao¹, Vilmar Æsøy² and Anne Stene¹

¹Faculty of Life Sciences

²Faculty of Maritime Technology and Operations

Aalesund University College

N-6025, Aalesund, Norway

E-mail: {yaca, aste, ve}@hials.no

KEYWORDS

Ballast water, micro organisms, heat treatment,

ABSTRACT

Ballast water contains a variety of organisms including bacteria, viruses and the adult and larval stages of the many marine and coastal plants and animals. As such, it poses serious ecological, economic and health problems and has serious negative effects on the global environment. This paper presents a new efficient ballast water analysis and heat treatment system using waste heat recovery system on ships. The project demonstrates laboratory methods to verify killing efficiency of micro-organisms in sea water exposed to heat treatment over a short period of time. Heating times were varied in a range from 20 seconds to 3 minutes. The micro-organisms were measured using a flow cytometry instrument and fluorescence microscopy to detect living and dead organisms in untreated and treated water. Based on the biological analysis, a related heat treatment simulation was carried out to confirm the control method.

1. INTRODCUTION

Water has been used as ballast to stabilise vessels at sea for over a hundred years. Ballast water is pumped inside vessels to maintain the correct operating conditions throughout a voyage. This is essential for safe and efficient modern shipping operations. This practice reduces stress on the hull, provides transverse stability, improves propulsion and manoeuvrability, and compensates for weight loss due to fuel and water consumption. Shipping moves over 80% of the world's commodities and transfers approximately three to five billion tonnes of ballast water internationally each year. Ballast water is essential to the safe and efficient operation of modern shipping, providing balance and stability to unladen ships [1].

Ballast water discharged by ships can have a negative impact on the marine environment. Cruise ships, large tankers, and bulk cargo carriers use a large amount of ballast water, which is often pumped-in from one region's coastal waters after waste-water has been

discharged or cargo has been unloaded. The ballast water is then discharged at the next port, or wherever more cargo is loaded. Ballast water discharge typically contains a variety of biological materials, including plants, animals, viruses and bacteria. These materials often include non-native, exotic species, which can cause extensive ecological and economic damage to aquatic ecosystems. Therefore, ballast water poses serious ecological, economic and health problems. Transferred species may survive and establish a reproductive population in the host environment, becoming invasive, out-competing native species and multiplying into pest proportions.

The purpose of this project is to develop methods for verification of treatment methods to be used in designing new ballast water handling systems at Aalesund University College. Our research is all based on the International Maritime Organization (IMO) Ballast Water Management Convention 2004, which will come into force in 2013 / 2014 after it has been ratified by nations representing more than 35% of the world's Gross Tonnage.

Both Flow cytometry and fluorescence microscopy were used and proven to be efficient instruments in the verification process. Furthermore, the biological tests indicated that the high temperature (80 °C) might be required in order to ensure efficient killing of micro-organisms (under ten microns) within 60 seconds of heating time. The larger organisms (phytoplankton and zoo-plankton) are almost killed at lower temperatures. After that, the water treatment process system was modelled using Bond Graph and 20SIM software. The objectives of the process simulation model were first of all to perform design optimisation on the different components in the system and second, to further simulate the dynamics in order to implement a proper control loop. At last, the conclusions and future work are given.

2. RELATED WORK

Scientists first recognised the signs and effects of the introduction of an alien species after a mass introduction

of the Asian phytoplankton algae *Odontella* (*Biddulphia sinensis*) in the North Sea in 1903. In the late 1980s, Canada and Australia were among countries experiencing particular problems with invasive species, and they brought their concerns to the attention of the International Maritime Organization's Marine Environment Protection Committee (MEPC) [2]. The problem of invasive species in ships' ballast water is now worsening. This is largely due to the expanded trade and traffic volume over the last few decades and since the volumes of seaborne trade continue to increase, the problem may not have reached its peak yet.

The last few decades have seen considerable progress in research on ballast water treatment and management. There is an urgent need in the shipping industry for the development of cost-effective and environmentally friendly Ballast Water Management Systems (BWMSs). According to the Ballast Water Convention, the International Maritime Organization (IMO) has set the Ballast Water Exchange Standard, D1 and the Ballast Water Performance Standard (BWPS), D2 (Table 1). The IMO Convention sets discharge limits on densities of live organisms based on the organism's size class. Organisms with a size that is greater than or equal to 50 μm mostly represent zoo-plankton, and organisms that are at least 10 μm but are smaller than 50 μm are mostly comprised of phytoplankton. Using zoo-plankton and phytoplankton as categorisation groups allows for a broad and important comparison of the results obtained in relation to the new standard [3, 4].

Table 1. The IMO ballast water performance standards (D-2)

Organism category	Standard
Organism size $\geq 50 \mu\text{m}$	< 10 viable organisms/mL
$10 \mu\text{m} \leq$ Organism size < $50 \mu\text{m}$	< 10 viable organisms /mL
Organism size < $10 \mu\text{m}$ (including the following items)	
Toxicogenic <i>Vibrio cholerae</i>	< 1 cfu/100 mL
<i>Escherichia coli</i>	< 250 cfu/100 mL
Intestinal Enterococci	< 100 cfu/100 mL

Many technologies have been under development during the negotiations at IMO, but it is difficult to compare the efficiency of treatments at removing organisms as, until the convention was adopted, there were no set standards.

The heat treatment of ballast water has been widely advocated as a possible treatment regime based on theoretical and laboratory/small scale trials. Various methods of heating the ballast water on board vessels have been previously used. The length of time the water was heated varied from 20 h at temperatures in excess of 35 $^{\circ}\text{C}$, 15 h at 42 $^{\circ}\text{C}$ and 80 h at more than 30 $^{\circ}\text{C}$ [5, 7].

Previous experiments carried out on board vessels have achieved a 90 – 100% reduction of the phytoplankton and zoo-plankton by using waste engine heat to treat the ballast at 35 – 38 $^{\circ}\text{C}$ for 20 h [6] and a 100% kill rate of zoo-plankton by heating the ballast water to 38 $^{\circ}\text{C}$ for 12 h [5, 7]. Instant exposures at high temperatures (40 – 65 $^{\circ}\text{C}$) have already been tested in the laboratory with successful results for phytoplankton and zoo-plankton [8, 9, 10].

This pre-project dealt with the application of this high temperature treatment under operational conditions. The aim was to assess the extent to which this method was able to treat the organisms smaller than 50 μm (phytoplankton and bacteria) in the ballast water.

3. TEST METHODS AND EQUIPMENT

3.1 Test equipment

Freshly collected sea water from "Åsefjord" containing natural strains of phytoplankton and the culture laboratory strain of *Escherichia coli* was prepared. The tests were performed at 51.8 $^{\circ}\text{C}$, 74 $^{\circ}\text{C}$, and 81 $^{\circ}\text{C}$ water bath for holding time of 10, 20, 30, 40, 50, 60, 90, 120 and 180 seconds. After treatment, the heated test tubes were directly removed from hot bath and placed into an ice bath to cool down for analysis. 100 $^{\circ}\text{C}$ treatment was used as a positive control. Test samples before and after treatment were then compared.

The viability of phytoplankton was measured by staining with SYTOX Green and carrying out tests involving flow cytometry assay and fluorescence microscopy. The viability of E-coli cells was measured by staining with SYBR Green, SYTO 9, or SYTOX Green, together with propidium iodide (PI), and tested by flow cytometry assay and fluorescence microscopy. Temperature of the water bath (Heto, Birkerød Denmark) at different levels: 22 $^{\circ}\text{C}$, 37 $^{\circ}\text{C}$, 35.5 $^{\circ}\text{C}$, 51.8 $^{\circ}\text{C}$, 69 $^{\circ}\text{C}$, 74 $^{\circ}\text{C}$ and 81 $^{\circ}\text{C}$.

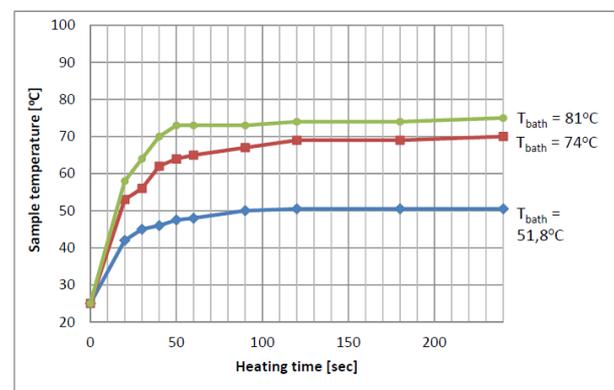


Figure 1: Temperature of 4 ml of water in Kimax glass tubes (10ml) in water bath at 51.8 $^{\circ}\text{C}$, 74 $^{\circ}\text{C}$ and 81 $^{\circ}\text{C}$.

Different test tubes were tried in the evaluation of heating efficiency by measuring temperature rise during heating in the water bath. These tubes and conditions

are: Nunc polypropylene tube (2 ml) with 2 ml of water, Kimax glass test tubes with screw caps (10 ml) with 10 ml of water, and Conical polypropylene centrifuge Tubes (15 ml) with 15 ml of water.

Results in our tests clearly show that Kimax glass test tubes provide the most efficient heating. As such, these are chosen for the further experiments. Temperature control in water bath was tested using 4 ml of water in 10 ml Kimax glass tubes measuring temperature rise in sample water. Results (Figure 1) show that the maximum temperature is reached after approximately 1 minute.

3.2 Organisms of different sizes

3.2.1 Organisms of size range between 10-50 μm

Organisms whose size are greater than or equal to 50 μm mostly represent zoo-plankton, and organisms whose size is less than 50 μm but more than 10 μm are mostly comprised of phytoplankton. Using zoo-plankton and phytoplankton as categorisation groups allows for a broad and important comparison of the results obtained in relation to the new standard.

Organism viability is not easily detected by a single morphological, physiological, or genetic parameter, making it advantageous to use more than one approach. Even procedures recommended in the protocol for land-based verification testing have practical limitations because of time constraints. There are no standard methods for readily and reliably discerning live and dead organisms of the 10-50 μm group. Flow cytometry can provide a rapid automated assessment of phytoplankton assemblages, especially since several recent advances have mitigated mechanical problems, like clogging, that have hampered its use in phytoplankton research in the past.

3.2.3 Micro Organisms small than 10 μm

Escherichia coli (*E. coli*) are 2-3 μm long rod-shaped bacteria, with a selectively permeable cell membrane and DNA held in a nucleoid area. Current best practice detection techniques for the viability of *E. coli* are based upon cell culture requiring 18 to 24 hours for a result.

Hopefully the device will be able to produce a quantitative result in less than an hour. Flow cytometry is rapid, easy, and sensitive for live/dead bacteria counting. Knowledge of the living/non-living and active/inactive states of cell populations is fundamental in understanding the role and importance of micro-organisms in natural ecosystems. Many approaches are based on membrane integrity, such as the Live/Dead kits (e.g. the LIVE/DEAD BacLight bacterial viability kit from Molecular Probes), a propidium iodide (PI) based assessment of dead cells. Usually, a combination of SYBR Green dyes or SYTO 9 and PI is widely employed for analysing live/dead bacteria numbers.

3.2.3 Calibration of flow cytometry size

The Flow Cytometry Size Calibration Kit has non-fluorescent particle-size calibration standards that provide a simple, accurate way to determine cell sizes by flow cytometry. The kit contains six suspensions of highly uniform polystyrene micro-spheres with the following diameters: 1.0 μm , 2.0 μm , 4.0 μm , 6.0 μm , 10.0 μm and 15.0 μm . The size of the plankton and bacteria were determined by comparison to standardised beads (10 and 50 μm).

4. WATER HEATING TREATMENT RESULTS

4.1 Water samples heated in a 74 °C water bath (flow cytometry results)

Fluorescence of SYTOX Green, a dye that only penetrates damaged cell membranes, and autofluorescence were observed simultaneously, allowing the discrimination of live and dead cells [12]. Freshly collected sea water samples were pre-filtered using a sieve made of a HYDRO-BIOS 50 μm mesh net. The 50 μm pre-filtered samples were for further treatment and analysis.

Before and after exposure to different length of heating, the 50 μm pre-filtered sea water samples were stained with SYBR and analysed on a flow cytometer (Figure 2). As shown in Figure 3, Natural red autofluorescence (red, FL-3H) from chlorophyll could identify viable phytoplankton cells (P1). Bright green fluorescence was observed in the samples that had been stained with SYTOX Green. These samples contained heated cells. After 20 seconds of treatment in a 74 °C water bath, 96% of the phytoplankton was killed (P2). After being heated for 180 seconds, no viable cells (P1) were detectable.

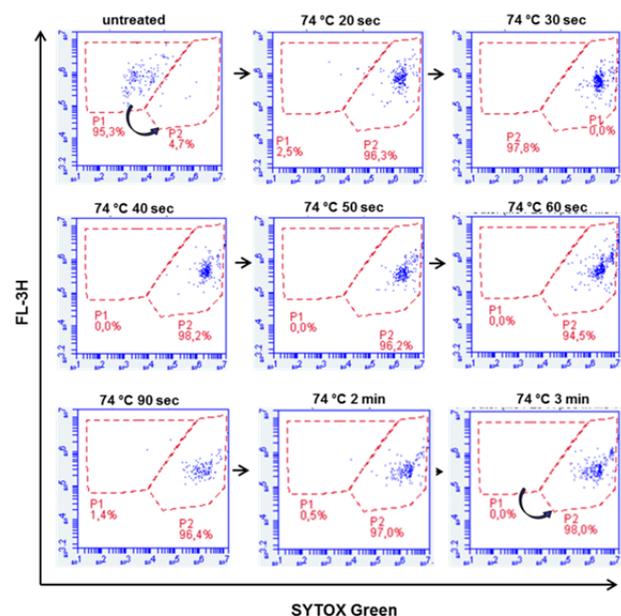


Figure 2: Flow cytometry results of 74 °C treated samples. P1: viable cells, P2: dead cells.

4.2 Comparison of 74 °C and 51.8 °C water bath

The 50 µm pre-filtered sea water samples were treated in either a 74 °C or a 51.8 °C water bath, then stained with SYBR and analysed on a flow cytometer, as shown in Figure 3. After three minutes of heating in water bath, the 51.8 °C treated group had still 1.2 % viable cells, while none were present in the 74 °C treated group.

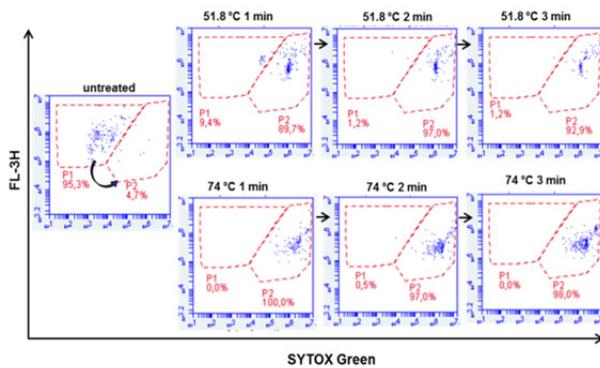


Figure 3: Treatment efficiency of a 74 °C and a 51.8 °C water bath. P1: viable cells; P2: dead cells.

4.3 Fluorescence microscopy test

Figure 4 shows that natural red autofluorescence from chlorophyll could identify viable phytoplankton cells (upper). Bright green fluorescence was observed in the samples that had been stained with SYTOX Green. These samples contained heated cells (lower). In the figure, the red part represents viable cells, and green part represents dead cells.

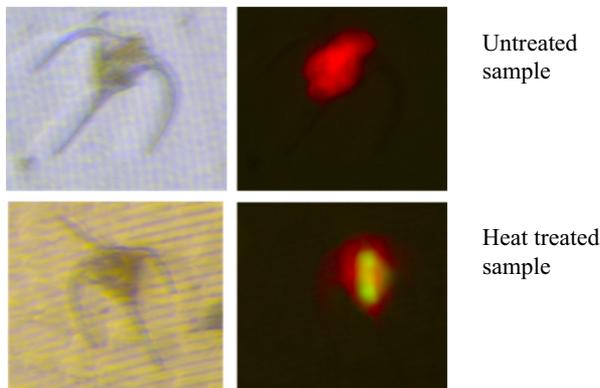


Figure 4: Natural red autofluorescence from chlorophyll.

4.4 E. coli heated in 80°C water bath

Flow cytometric analyses of E. coli before and after being heated are shown in Figure 5.

Before and after exposure to different amounts of heating, bacterial cell samples were stained either with a mixture of SYBR Green plus PI or with only SYBR Green and analysed on a flow cytometer. After 20 seconds of treatment in an 80 °C water bath, staining with SYBR Green and PI showed intermediate states.

After being heated for 90 seconds, all cells were PI positive. The green colour show viable cells, while the red indicates dead cells.

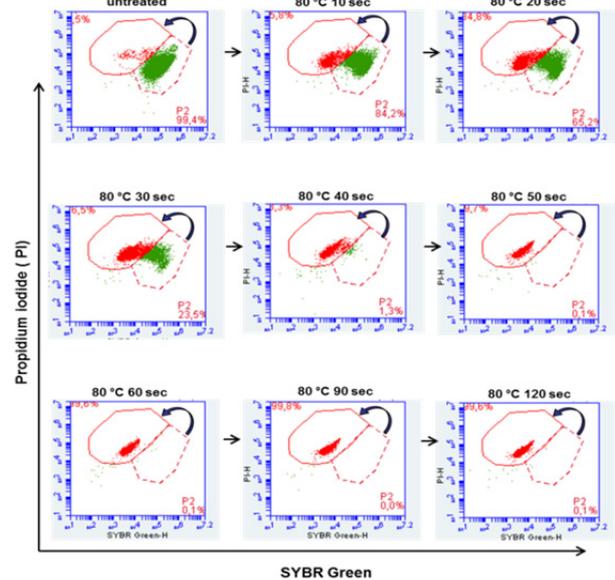


Figure 5: Flow cytometric analysis of E. coli before and after heat treatment. Green: viable cells, Red: dead cells.

4.5 Three methods for E. coli viability detection

Flow cytometric analysis of E. coli. Bacterial cells were heated at 100 °C for one minute or in an 80 °C water bath for two minutes. Untreated and treated bacterial cell samples were stained with a mixture of either SYBR Green (left), SYTO9 (middle) or SYTOX Green (right) plus PI and analysed on a flow cytometer. Results are plotted in Figure 6, where P1 (red) are dead cells P2 (green) are viable cells.

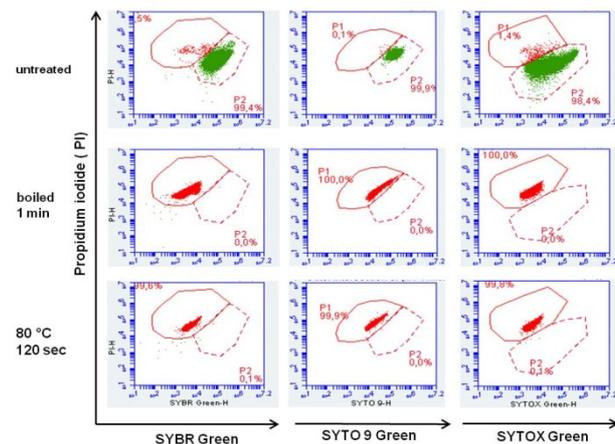


Figure 6: Flow cytometric analysis of E. coli bacterial cells heated at 100 °C for one minute or in 80 °C for two minutes. P1(red): dead cells P2 (green): viable cells.

Further fluorescence microscopy analysis of E. coli was performed on the same samples. Results are shown in Figure 7. This project has demonstrated laboratory methods to verify killing efficiency of micro-organisms

in sea water exposed to short time heat treatment. Heating times were varied in the range from twenty seconds to three minutes. Flow cytometry assay and fluorescence microscopy are used to measure living and dead organisms in untreated and treated water.

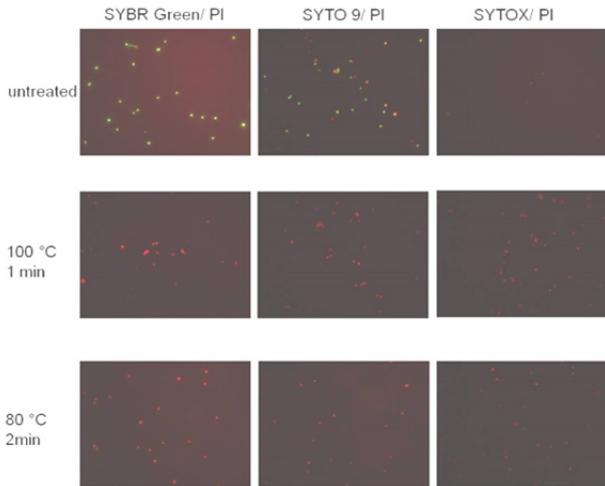


Figure 7: Fluorescence microscopy analysis of E. coli.

5. BALLAST WATER TREATMENT SIMULATIONS

In this part of the project we will develop simulation models for a ballast water treatment system that employs a waste heat recovery system. The simulator should allow for the optimisation of heat use by way of an efficient combination of heat exchangers, reactors and a control system. The aim is to assess the extent to which this method is able to treat the organisms under $50 \mu\text{m}$ (phytoplankton and bacteria) in ballast water within a heating time of one minute.

The schematic of the water treatment process system is shown in figure 8. The main components of the system are the *heat recovery unit* and the *reactor* where bio-treatment follows a temperature-time history.

The system is modelled using Bond Graph method [13, 14, 15] and 20SIM software [16]. The overall system model is shown in Figure 9 and the heat recovery unit is shown in Figure 10. The Bond Graph is a unified

modelling tool for multi domain systems where energy flow (power) and preservation of energy and mass are the common variables. The basic model consists of energy storing elements (C-elements and I-elements) and energy transferring or dissipative elements (R-elements). In figure 10 the R-elements represent the convective and conductive heat flows while the C-elements are the heat capacitance in each model segment. The I-elements represent the hydraulic inertia in the fluid flow. An icon-based object-oriented modelling interface is provided by 20SIM, where the overall system layer is shown in Figure 9.

The objectives of the process simulation model are to optimally design the different components in the system and to further simulate the dynamics in order to implement a proper control loop. The purpose of the treatment system is to heat and keep the temperature at $T_{\text{reactor}} = 80 \text{ }^\circ\text{C}$ for 30 seconds. The water is then cooled down in the recovery unit in order to save energy. The objective is to minimise the need for boost heating in the super-heater.

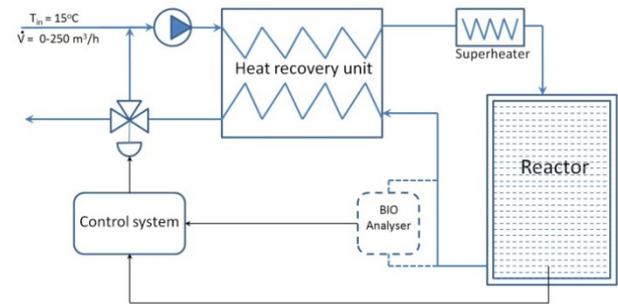


Figure 8: Water treatment system.

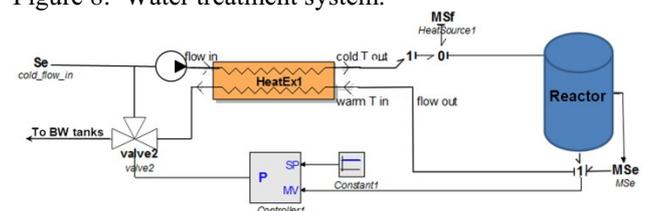


Figure 9: Treatment system simulation model.

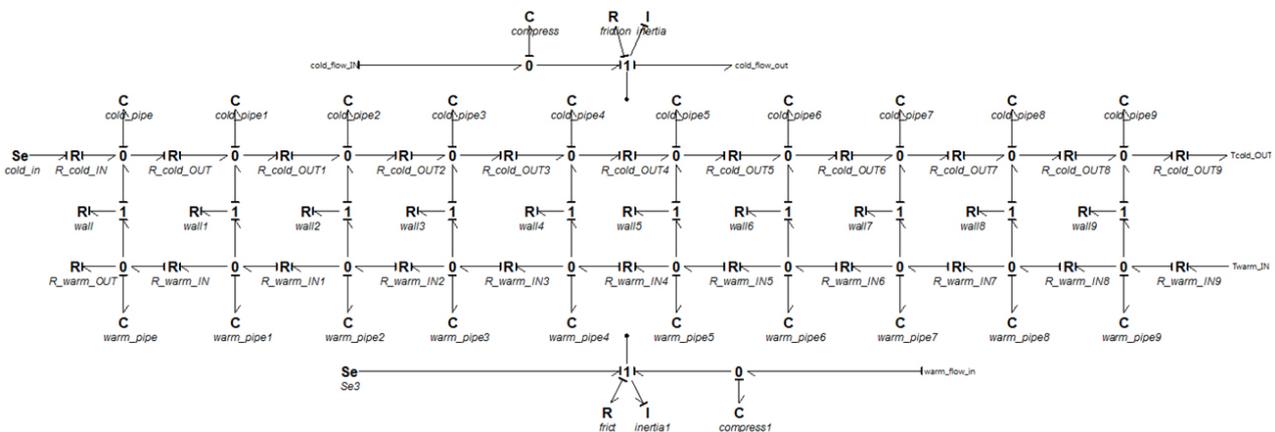


Figure 10: Heat recovery unit model.

In the first part of the simulation study, the initial heating and cooling process was simulated. The results showed that the system reached the required temperatures and initial heating time as required. Figure 11 shows the temperature-time history for the reactor and inlet/outlet temperatures for the heat recovery unit simulate a start and stop cycle. Further simulations included a control loop to secure optimal flow and power control for a continuous process. Figure 12 shows the temperature history for the water flowing through the continuous process simulating different flow control. The next steps will be to implement a Bio-analyser and to model the bio-reactor function.

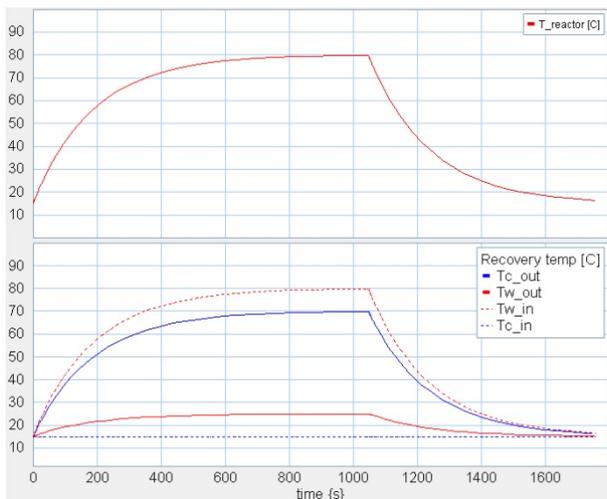


Figure 11: Simulated heating and cooling process.

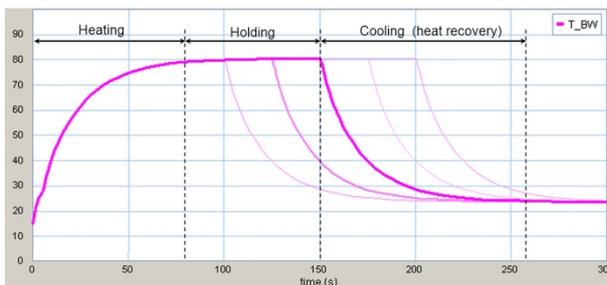


Figure 12: Water temperature history simulation using different flow control parameters.

6. CONCLUSIONS AND FUTURE WORK

Ballast water treatment is still a very challenging area. Many technologies have been under development during negotiations at IMO, but it has been difficult to compare the efficiency of treatments at removing organisms as, until the convention was adopted, there were no set standards. The first step in our project is to investigate the current ballast water treatment situation, and to focus on the methods. Therefore, more extended tests are needed to evaluate the efficiency of the treatment method.

From a biological analysis viewpoint, both flow

cytometry and fluorescence microscopy have proven to be efficient instruments in the verification process. The results from the preliminary experiments indicated that heat treatment is efficient in reducing the viable phytoplankton in natural sea water and laboratory cultures of *Escherichia coli*. The killing efficiency depends on the heating temperature and the holding time. This result will be further verified using the classic membrane filtration method in the future work.

Furthermore, the preliminary results indicated that a high temperature (80 °C) is probably required for the bacteria under 10 microns in size (such as *E. coli*), in order to guarantee efficient killing within 60 seconds of heating time. The larger organisms (phytoplankton and zoo-plankton) are killed at lower temperatures.

The ballast water treatment simulation verifies the possibility of using waste heat recovery systems on board ships to deal with the problem. The final goal of the project is to develop methods and equipment for verification and energy optimisation in design of new ballast water treatment and management systems.

The proposed project is still the first step of a long-term project. We will have the following work including developing biological methods for verification of ballast water treatment, developing a ballast water management simulator using waste heat recovery, and proposing the treatment machinery system for real applications.

ACKNOWLEDGMENTS

The project is supported by a VRI-project in Norway. The research work is based on the close cooperation with ULMATEC – Pyro. The authors would like to thank for the contribution from Mr. Yue Li for his support on the 20-sim modelling.

REFERENCES

- [1] <http://globallast.imo.org/index.asp?page=problem.htm&menu=true>
- [2] <http://www.imo.org/OurWork/Environment/BallastWaterManagement/Pages/Default.aspx>
- [3] Anonymous, Guidelines for approval of ballast water management systems (G8). Annex3 Resolution MEPC.125(53), Annex, Parts 1,2,3 and 4., 2005.
- [4] PhD, M.J.W.V., Final report of the land based testing of the BalPure® BWT System. Royal Netherlands Institute for Sea Research, 2009.
- [5] Quilez-Badia, G., et al., On board short-time high temperature heat treatment of ballast water: a field trial under operational conditions. *Mar Pollut Bull*, 2008. 56(1): p. 127-35.
- [6] G. Rigby, G.M.H., C. Sutton, Novel ballast water heating technique offers cost-effective treatment to reduce the risk of global transport of harmful marine organisms. *Marine Ecology Progress Series*, 1999. 191: p. 289-293.
- [7] Mountfort, D.O., Dodgshun, T., Taylor, M., , Ballast Water Treatment by Heat - New Zealand Laboratory and Shipboard trials. In: 1st International Ballast Water Treatment R&D Symposium, 26-27 March, 2001, No. 5.

IMO, London, pp. 45-50. 2001.

- [8] McCollin, T.A., Shanks, A., Biological Assessment. Phytoplankton Results. DTR-3.7.2-FRS-06.03. MARTOB - On Board Treatment of Ballast Water (Technologies Development and Applications) and Application of Lowsulphur Marine Fuel, Newcastle upon Tyne, UK, 27th June, 2003, 87pp. 2003.
- [9] Quílez-Badia, G., Gill, M.E., Frid, C.L.J., Biological Assessment. Zooplankton Results. DTR-3.7.1-UNEW-08.03. MARTOB - On Board Treatment of Ballast Water (Technologies Development and Applications) and Application of Lowsulphur Marine Fuel, 29th August 2003, Newcastle upon Tyne, UK, 39pp. 2003.
- [10] Euan D. Reavie, A.A.C., and Lisa E. Allinger Assessing Ballast Water Treatments: Evaluation of Viability Methods for Ambient Freshwater Microplankton Assemblages. *Journal of Great Lakes Research* 2010. 36(3): p. 540-547.
- [11] Berney, M., H.U. Weilenmann, and T. Egli, Flow-cytometric study of vital cellular functions in *Escherichia coli* during solar disinfection (SODIS). *Microbiology*, 2006. 152(Pt 6): pp. 1719-29.
- [12] Masanori Sato, Y.M., Mika Mizusawa, Hitoshi Iwahashi, and Shu-ichi Oka, A Simple and Rapid Dual-fluorescence Viability Assay for Microalgae. *Microbiology and Culture Collections*, 2004. 20(2): p. 1342-4041.
- [13] Pedersen, E., Modelling multicomponent two-phase thermodynamic systems using pseudo-bond graphs, 2001 International Conference on Bond Graph Modelling and Simulation (ICBGM'01); Nov. 2001.
- [14] Karnopp, D.C., Margolis, D.L., & Rosenberg, R.C., *System Dynamics: A Unified Approach*. John Wiley & Sons, Inc., second edition, 1990.
- [15] Thoma, J.U., & Richter, D.B., Simulation of Fluid Pipes in Hydrostatic Circuits Using Modal and Segmented Methods. *Transactions of The Society of Computer Simulation*, Vol. 3, (no. 4):pp.337-349, October 1986.
- [16] <http://www.20sim.com>

AUTHOR BIOGRAPHIES

Yanran Cao received her M.D. degree in 2004 from the Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS & PUMC) in China. From 2004 to 2011, she worked as a post-doctoral research fellow in one group of tumour immunology, Department of Oncology/Haematology, University Medical Centre Hamburg-Eppendorf (UKE). In 2013, she received her Ph.D in Biology, Department of Biology, University of Hamburg. Since January 2012, she has worked as a researcher at the Faculty of Life Science at Aalesund University College.

Vilmar Æsøy received his Ph.D. in Mechanical Engineering in 1996 from Norwegian University of Science and Technology. From 1997 to 2002 he worked as a researcher in Aker Maritime and R&D manager in Rolls-Royce Marine As. Since 2002 he has been working as an associate professor at Aalesund University College.

Anne Stene received her Dr. Scient. in 1998 as judged by a selection committee. In 2013, she received a Ph.D in epidemiology from the Norwegian School of Veterinary Science. Since 1998 she has worked with education, research, administration and consultancy at Aalesund University College regarding marine ecology, fishery and aquaculture. She has also worked in the Norwegian Directorate of Fishery and in the Norwegian Ministry of Fishery and Coastal affairs.

SIMULATION CHALLENGES IN THE DEVELOPMENT PROCESS OF A COMPLEX PRODUCT: DESIGN OF VIRTUAL ELECTRIC SPORTS CAR

Eszter Varga

Attila Piros

Balázs Vidovics

Department of Machine and Product Design

Budapest University of Technology and Economics

H-1111 Műegyetem rkp. 3-9. Budapest, Hungary

E-mail: eszter.varga@gt3.bme.hu

KEYWORDS

Product Development, Virtual Simulation, Virtual Product, Digital Mock-Up, Higher Education.

ABSTRACT

The project being introduced is a simulation of an industrial Product Development project as a large scale student project in a university setting. The scientific goal with the project is to develop and test different management, design and engineering simulations, the practical goal however is to provide a real-like project environment for the students, which fit to academic conditions most perfectly. The target of the development process is to design a fully electric sports car. The outcome of the project would be a detailed Digital Mock-Up, on which Virtual Simulations will be carried out. In the paper the background necessary for establishing the process simulation will be presented, besides the successful realization will be proven by showing the recent achievements from the project. The Digital Mock-Up is being developed by applying the Concurrent Engineering approach, the development process is closely supported by project management activities. The necessary virtual test will be carried out by university students on the specific parts of the sports car following the protocols being used in industry. The realization of a simulation project of such a complexity provides great opportunity for both students and teachers/researchers to gain invaluable knowledge and scientifically relevant findings.

INTRODUCTION

It is above all dispute that the role of highly detailed Virtual Simulation (VS) methods and tools in the Product Development (PD) process is important in industrial environment. Companies face increasing number and increasing complexity of challenges throughout the development process. Products reach higher and higher complexity and in parallel the time to market period decreases very fast (Becker et al. 2005). With the application of VS companies handle those two factors reasonably well. VS models and tests the behaviour of the product or a component from a specific perspective by simulating realistic environment, therefore enables companies to reduce design time and

design risks. The rapid development of Information Technologies (IT) continuously provides new opportunities for emergent and developing VS applications ranging from different Finite Element Analyses (FEA) to maintenance simulation (Alabdulkarim 2011).

Nowadays no company could introduce a competitive product without the utilization of upper mentioned virtual technologies; and the same rule applies to Higher Education (HE) as well. However, universities do not produce competitive products but disseminate competitive knowledge to foster the creation of competitive products. This paper will focus on technical HE, with a closer look on the field of engineering design. Engineering design students have to have up-to-date knowledge in their specialization, which is impossible to get acquired without getting familiar with the latest VS technologies. To achieve this goal, HE institutions may have two options.

Many universities in Hungarian HE have the opportunity to collaborate with industrial companies in industry-academia PD projects. In this frame students meet the industrial procedures and protocols and work as 'subcontractors' in a project. This type of collaboration is not uncommon, however it has its pitfalls, e.g. the mismatch between the company and the university expectations, the different time-frames and time-scales of the project and the academic calendar, etc. The major characteristic of an industry-academia project is that the design assignment, most of the input requirements and specifications come from the company side, and the project manager is at the company, too. While the decisions are mainly made inside the company, the project management workload is on the HE institution. A generic model of a VS oriented PD process is shown on Figure 1.

For this reason it is assumed that the simulation of industrial PD processes within the frame of educational projects would fit the HE institutes' goals and resources better. This kind of complex VS was missing from the palette of the Hungarian HE, so the initial idea was to establish a *virtual simulation of an industrial project*. This was meant to model all the main characteristics of

an industrial PD process in one single educational industry-independent academic project, which literally meant to exclude any company involvement on the input side.

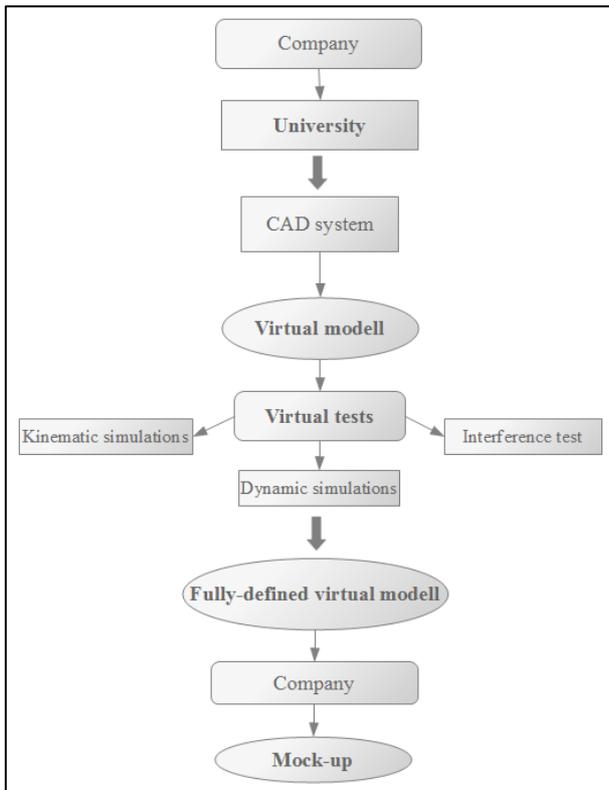


Figure 1: The Generic Model of a VS oriented PD Process in HE Environment

MOTIVATION AND CHALLENGES

The goal of the VS in question is to create a complex virtual product, the output from the PD process will be a Digital Mock-Up (DMU). This DMU is being created according to the Concurrent Engineering (CE) philosophy. CE is a commonly accepted design method in industry for decreasing the time to market of a product (Clark and Fujimoto 1991). The VS is being continuously supported by a variety of project management tools. These tools ease the handling of complex, multi-participant design projects (Cho and Eppinger 2005).

Planning and scheduling of the PD process is always hard because the output and the run-off paths are affected by numerous factors and therefore the risks are high (Szélig et al. 2011). The biggest challenge for the teachers involved in the project is the set-up of the theoretical and technical background for the simulation of the industrial PD process. Teachers have to model the project run-off in advance and identify the possible risks, which calls for the application of different project management tools at the teachers' side. Parallel to that, the technical background of the VS must be set up by installing a database providing access for all participants

of the project. Students also need support in terms of software access and usage for the successful execution of the VSs. On Figure 2 the overall simulation environment of the project is presented.

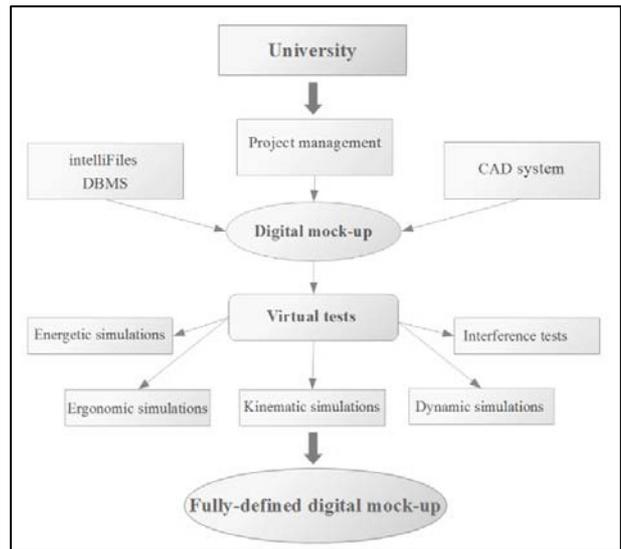


Figure 2: The Simulation Environment of the Project

This kind of design project is fruitful for both students and teachers. Students can extend their theoretical knowledge and apply it by participating in a project simulating a real one. Students become experienced with the work-flow where a number of engineers work on the same product in parallel, and where also the boundaries and the parameters are either pre-determined or related. The DMU provides a good opportunity for trying out the acquired knowledge on the VSs.

As it was described earlier, exclusively university students are involved in this project. Due to the complexity and cross-disciplinary character of the project, dynamic multi-disciplinary student teams were formed, where students with different specializations (engineering design, mechatronics, industrial design engineering, management studies) are working together. It is a highlighted aim in the future to involve the most students from different fields, which is made possible by the virtual characteristic of the project. The set-up of the students involved is illustrated on Figure 3. The teams are changing continuously and flexibly according to the project needs: as the different sub-systems are being developed or as more specific knowledge is necessary, etc.

Number of students	Mechanical Engineer	Industrial Design Engineer	Mechatronics Engineer	Engineering Manager
2013 I.semester	2	-	-	-
2013 summer	4	-	2	-
2013 II. semester	4	1	1	1
2014 I. semester	2	1	-	-

Figure 3: The Mix of Students by Specialization

The introduction of this PD simulation project into education means new challenges to the university and also provides invaluable experiences to the students. This kind of complex VS also has high scientific potential. The next section introduces the VS more in details.

THE OVERVIEW OF THE PROJECT

The subject of the simulation of the PD process is fully electric driven virtual sports car (Figure 4). For the first sight the fully electric propelled vehicle looks highly innovative but the idea itself dates back long in the history. The first fully electric vehicle was created by French G. Trouvé in 1881, four years before Carl Benz was to demonstrate the first operating internal combustion engine vehicle (both were tricycles) (Westbrook 2001). Nowadays the pure electric drive comes to the front again thanks to the increasing weights of the environmental aspects. It is common sense, local pollution is not emitted by solely electric driven vehicles. Furthermore – in comparison –, the drive-train of the electric sports car is much simpler than the drive-train of an internal combustion car, for e.g. it requires less moving parts and auxiliary components.

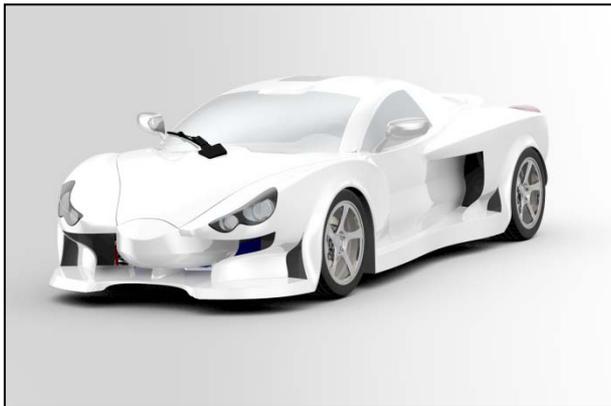


Figure 4: The Electric-Driven Sports Car

A uniquely electric powered vehicle has its weak points as well. Most of the difficulties are caused by the batteries because due to their size and weight, although this is the most intensive area of research. Currently the commercial battery technology and infrastructure are ‘half-baked’; in practice, a short range of distance and on top of that, rare charging stations and long recharge times limit the usage against the internal combustion (and hybrid) technology.

Independently of these disadvantages electric vehicles have great perspective, so leading car manufacturers cannot afford to ignore this type of cars. This area is interesting and – what is more important – really a rewarding area for HE and research institutions. These reasons lead us to choose this subject for the VS project.

Input Parameters

The car design has many fixed parameters and boundary conditions. As a first step, the two most influencing design tasks, namely the conceptual design of the drive-train and the car exterior design have been completed by two students in their major thesis works. This can be considered as the preliminary design of the vehicle, which then provided the basis for designing the upcoming sub-systems.

At the beginning of the project the following initial parameters have been defined. The car is a two-seater and can transport a driver and one passenger. The sporty style of the car suggests a racing application but this car is intended to be a typical second car of the family complying with the rules of the road. The core of the chassis is a carbon monocoque cell extended with two auxiliary aluminium frames. Similarly to the monocoque body the material of the outer shell is also carbon composite. The driving power is provided by two YASA 400 electric motors, each 400 Nm of maximum output torque, independently built in the rear. There are two mechanical gearboxes attached to the electric motors. The 200 km range with an average speed of 150 km/h with a single recharge was an also important initial condition. Lithium-Polymer batteries with 85 kWh capacity was chosen to meet the previous requirement. The next chapter describes the realization of this project.

SIMULATION OF INDUSTRIAL ENVIRONMENT

A project with such a high complexity and long duration requires special preparation before its start. On the one hand those are related to project planning, on the other hand those are concerning to the technical background.

Scheduling

The scheduling of the design process needs careful preparation. The final success of the project is significantly influenced by the proper planning of these design steps.

Since the PD is simulated as a project the preparations regarding the work-flow was the project management is based on traditional methods. After the specification of the project goal the network diagram was developed. The first step was to analyze the system of the sports car and to identify the design tasks in order to create the Work Breakdown Structure WBS) (Haugan 2001). A network based technique was applied to schedule the project. The Critical Path Method (Kelley and Walker 1959) was combined with the PERT evaluation (Fazar 1959) The traditional time estimation was replaced with a fuzzy based evaluation method (Piros and Veres 2013) in the scheduling procedure. Based on these techniques the whole project network diagram was created including the Critical Path (CP) (Figure 5).

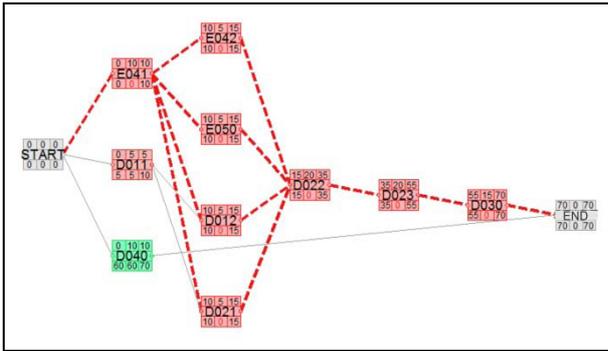


Figure 5: The PERT Network Diagram with the CP

Resources (students working in semester projects, major and final projects, etc.) then could be allocated to the project accordingly, and the participating students are possible to be monitored to collect continuous and accurate feedback about the status of the project. This data is post-processed for further research purposes to assist in the development of a new mathematical model to model and estimate risk of design processes.

Design Approach

During the preparatory phase of the project the formation of the technical background plays just as important role as project scheduling. The first step in the realization of the virtual model of the sports car was the creation of the product structure which guarantees the proper location of each component in such a complex system. During the realization of the product structure the selected design philosophy must be taken into consideration. This selected design approach was the Top-Down Design (TDD) approach which comes with many advantages. The TDD method controls the whole design procedure from the top level of the design structure, and suits well to the custom-designed systems. When applying the TDD the controlling element is the overall concept called the design skeleton, which is the basis of the whole virtual model. The model structure is broken down into sub-assemblies then the individual components are being designed according to the concept. The TDD approach generally requires a low amount of input information, and all modifications are initiated on the conceptual level. This method fits well to CE providing the opportunity to design the different sub-assemblies in parallel (Misra 2008). In the meantime, in line with the definition of the product structure the creation of the related data structure is also required according to the hierarchy levels of the system to be able to manage and store the files of the assemblies, sub-assemblies and individual components. In order to achieve this a novel generic identification system was developed which is assumed to be applicable in future projects as well.

IT Support

Once the model structure is ready the concepts on the specific assembly levels could be developed, which

obviously require a 3D CAD system. Budapest University of Technology and Economy (BME) is in the lucky position to have access to many makes of CAD software and maintain good relationship with the resellers. Thanks to that the project is supplied with educational software licenses. As the main computer support tool in the design process the monolithic PTC Creo 2.0 CAD environment is being used by the participants. This CAD system is widely used in industry, and has all the required features to realize the DMU and has sufficient modules to execute the different VSS.

In the case of complex technical projects, where numerous engineers collaborate the application of a Product Data Management (PDM) system is of essential importance. The PDM systems handle all the necessary data in design and manufacturing processes of the products (Crnkovic et al. 2003). These PDM systems are built on Data Bases (DB) which permanently store data in organized form. For the handling of the DBs the application of Data Base Management Systems (DBMS) is required. This will enable the users to store and handle data and assist them to evaluate, search, retrieve and track changes on the data (Ullman and Widom, 1997).

Generally a commercial DBMS system is applied to handle this kind of project data. In our case a self developed PDM system named intelliFiles (iF) was applied (Figure 6).

The iF has all of the important features of a commercial PDM/PLM system as listed in the following:

- functions to store and manage any kind of project data,
- query ability,
- data change management with logging,
- advanced backup and replication,
- security functions with user privileges management,
- change and access logging,
- file/document life-cycle support with different roles of the users,
- CE support with semi automatic communication.

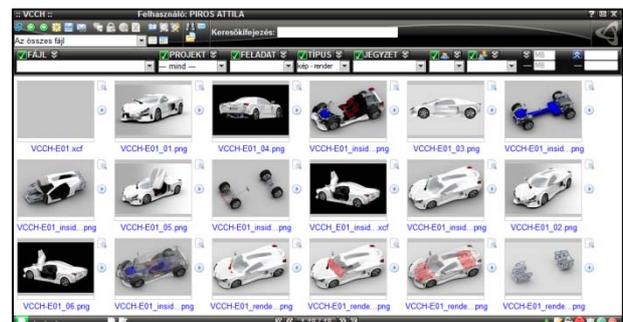


Figure 6: The intelliFiles window

These functions show that the handling of such a complicated project like the design of an electric sports car can be supported with iF. Most importantly it enables a number of students working in parallel on the model while product data is kept organized and managed.

The previous sections illustrate that both the theoretical and technical conditions of the project are sufficient and well simulate the processes and methods applied in industry, therefore authors argue that the key aspects of an industrial PD process has been successfully implemented in an academic environment through an autonomous design project.

THE DIGITAL MOCK-UP

The ultimate target of the simulation project is the creation of a high complexity CAD model, a Digital Mock-Up, which is used in several fields of design and manufacturing. An active DMU has most of the properties of a physical product model, therefore it is a good subject of different VSs to substitute the physical tests.

Taking the design process into consideration – as already mentioned above – after the brief specification of the product the first steps of the conceptual phase were the creation of the exterior design (instantly modelled as a surface model in the and the CAD system) and the preliminary design of the drive chain.

Those two top-level systems determine the design since most of the car parameters are derived from these two main components. During the design of the DMU and its sub-systems all the participating students had to observe some highlighted rules.

Each part of the car has to meet the related standards, regulations and other applying rules. It was an important expectation towards participant students that all components have to be designed for manufacturability and in some specific cases the design of the manufacturing tool was also required. User requirements and consumer protection aspects have also been taken into consideration during the design of the concepts. Ergonomic studies have been executed on the components which are in interaction with the users. In these latter cases the design must meet the ergonomic requirements for the satisfaction of future owners' needs. There have been some examples when considerably different ones from the traditional solutions or rather innovative solutions emerged, then patent research were initiated.

Taking all those requirements into consideration in the DMU of the electric sports car the following components and sub-systems have already been developed and incorporated (see Figure 7 for visualization):

- exterior surfaces,
- drive-train,
- steering mechanism,
- brake system,
- monocoque central body part,
- auxiliary frames on at the front and rear,
- door concept,
- hood concept,
- rear-view mirror concept,
- windshield wiper,
- interior concept.

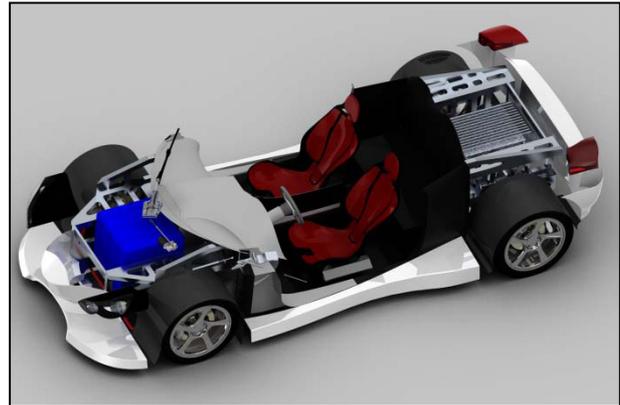


Figure 7: The Readiness of the Electric Sports Car

These tasks below are currently under preparation:

- detailed design of the car body shell elements,
- detailed design of the door and the side mirrors,
- detailed design of the interior of the car.

The listed tasks cover a working period of eight months and the high number and the complexity of the components well demonstrate the success of this project so far.

In this section the DMU was described as a product model structure, but on the other side of the application of the DMU there are the VSs executed by the students in relation with all components and sub-systems. A few of these simulations are introduced in the next section.

VIRTUAL TESTS ON THE DIGITAL MOCK-UP

Since the goal is to build a DMU, we aimed at carrying out all possible tests and simulations, the experimental application of the DMU can only be complete with the execution of the VSs. Some simulations were executed by the students based on their current knowledge in that specific field, but in some other cases students had to acquire new knowledge to succeed, or developed their competences by working with the DMU. The application of the learning-by-doing principle is found very useful according to the feedbacks from the students.

The first virtual test executed was the energy engineering simulation of the car. This VS was really essential in this special case of the electric sports car, as

numerous other parameters were derived from the results of this VS in question. The simulation itself was oriented to map and estimate the total energy consumption of the car. During the execution of this VS the necessary energy and torque to move the car were calculated. On the basis of the results a specific electric motor was later selected. The other important component which significantly determines the arrangement of the car's drive-train is the battery pack. The battery cells were also selected upon the results of the previous VS and then the arrangement of the battery pack was finalized. In the next steps the auxiliary electrical devices were selected e.g. the steering servo, air conditioning unit, windshield wiper drive and the lighting instruments. After the selection of these components the energy consumption of the car was further iterated and evaluated. There came an idea about the possible application of outer solar cells to supply the auxiliary devices. The visible solar cells also make the car looking much more environment friendly and this would definitely be an attractive feature for the future owners. The main output of this VS was the total energy balance of the electric sport car.

The following simulations were parallelly executed by the students based on the method of the Concurrent Engineering. A car has to be a subject of many different ergonomic simulations since many components of the car are directly in interaction with the users. A VS was created to evaluate to what extent the selected and proposed components comply with the related regulations. These regulations are derived from the anthropometric measures of the human body. The applied CAD system provides tools to execute these ergonomic VSs. The software offers computer human models (i.e. manikin models) with various percentile dimensions. The application of these models proves the compliance of the design with the broad range of the population. These VSs included the vision evaluation from inside the car (Figure 8), the positioning of the steering wheel and the seat, the positioning of the rear-view mirrors and finally the reach simulation of the operator instruments and buttons.

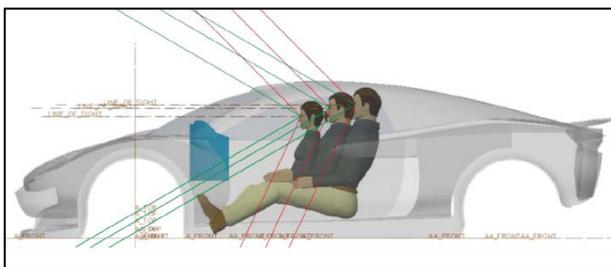


Figure 8: The Ergonomics Simulations: The Vision (Field of View) Evaluation

The simulations and tests commonly used in engineering design practice were applied in a wide range. Student teams carried out interference tests for instance. Furthermore, many kinematic and dynamic

simulations have been executed during the project, e.g. the dynamic simulation of the suspension components was of absolute necessity because of the high load of the built in parts. On top of these previously mentioned simulations many other finite element analyses were executed in the field of mechanical and heat transfer processes. The majority of the simulations focus on the critical components of the car. These virtual simulations have took out many future physical tests.

CONCLUSION AND FURTHER RESEARCH

The simulation of an industrial project has many useful aspects. Practical results like the experiences from the process simulation of the project or the DMU finally realized can be presented. The educational requirements are all fulfilled, as students get the access to the latest IT tools, they learn theory and practice (learning-by-doing) in good balance. The virtual simulation of an industrial project can always be tailored to the needs of the HE institute and the students.

However, BME is not only an educational institute but also a research university. This is the reason why this project is being approached from the scientific point of view. The realization of this project can also be considered as a scientific achievement since it required a detailed background research and on the fields where insufficient or even the lack of tools or methods were detected, bespoke or customized tools or methods have been developed for the purpose of the project. One great example would be the process management method developed, the other would be the DBMS; this special IT tool has been specially customized to suit the requirements of this VS. Besides, further scientific results created at the university were utilized or tested in the frame of this project. The participation in the concurrent scientific research is not only available for the university researchers, but many students are also involved to this project, e.g. students wrote their diploma projects about working on the virtual sports car project. Two research areas must be highlighted in this project.

One research area is related to the processing of the monitored data of the project execution. This research focuses on a mathematical model to simulate and track the special risks in the design process. The risk assessment is based on fuzzy method like published by Retter (2007). An average project execution is loaded with different risks which is particularly typical in the design processes (DashWu et al. 2010) therefore any new mathematical model oriented to handle these risks might be of great step ahead in this field. The other research focus is related to the optimization and automatic set-up of the side mirrors. This concept is highly innovative, so it is under consideration to be patented. These two examples already well represent the scientific potential of this VS. This VS will surely triggers other research projects in the future and might turn some student's interest towards a scientific carrier.

ACKNOWLEDGEMENTS

This research of Vidovics, B. underlying this paper was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 'National Excellence Program'.

REFERENCES

- Alabdulkarim, A.A.; Ball, P.D.; and Tiwari, A. 2011. "State of the Art of Simulation Applications in Maintenance Systems" In *Proceedings of the 44th CIRP Conference on Manufacturing Systems* (Madison, Wisconsin, USA, June 1-3). CIRP, Paris.
- Becker, M.C., Salvatore, P., and Zirpoli, F. 2005. "The impact of virtual simulation tools on problem-solving and new product development organization." *Research Policy*, 34, No.9, (Nov), 1305-1321
- Cho S.H. and Eppinger, S.D. 2005. „A simulation-Based Process Model for Managing Complex Design Projects” *IEEE Transactions on Engineering Management*, 52, No.3.
- Clark K. and Fujimoto T. 1991. „*Product Development Performance: Strategy, Organization, and Management in the World Auto Industry.*” Harvard Business School, Boston, MA.
- Crnkovic, I.; Asklund, U.; and Dahlqvist, A.P. 2003. *Implementing and Integrating Product Data Management and Software Configuration Management*. Artech House, Norwood.
- DashWu D.; Kefan X.; Gang C.; and Ping G. 2010. "A Risk Analysis Model in Concurrent Engineering Product Development" *Risk Analysis*, 30, No.9.
- Fazar, W. 1959. "Program evaluation and review technique", *The American Statistician* 13. No.2., 646-669.
- Haugan, G.T. 2001. *Effective Work Breakdown Structures*. Management Concepts, Vienna.
- Kelley, J. and Walker, M. 1959. "Critical-path planning and scheduling," In *Proceedings of the EJCC 1959*, 160-173.
- Misra, K.B. 2008. *Handbook of Performability Engineering*. Berlin, Springer.
- Piros A. and Veres G. 2013. "Fuzzy based method for project planning of the infrastructure design for the diagnostic in ITER" *Fusion Engineering and Design*. 88. 1183-1186.
- Retter Gy. 2007. "Kombinált fuzzy, neurális, genetikus rendszerek" Invest-Marketing Bt., BME-VET, Budapest.
- Szélig N.; Vidovics B.; and Bercsey T. 2011. "Time-estimation of design process based on patterns" *Periodica Polytechnica – Mechanical Engineering*. 54. No.1. 57-62.
- Ullman, J.D. and Widom, J. 1997. *A first course in database systems*. Prentice-Hall, Upper Saddle River.
- Westbrook, M.H. 2001. *The Electric Car: Development and Future of Battery, Hybrid and Fuel-cell Cars*. IEE, London.

AUTHOR BIOGRAPHIES



MS. ESZTER VARGA was born in Budapest, Hungary in 1991 and went to the Budapest University of Technology and Economics (BME), where she studied Industrial Design Engineering and took her bachelor degree in January 2014.

After that she started the Master Education in Mechanical Engineering. In 2012 she was researching the methods of reconstruction of virtual pieces, and nowadays she is studying project management, specializing in fuzzy based risk assessment. She also participates in the education activity of the Department of Machine and Product Design. She is the corresponding author of the paper, her e-mail address is: eszter.varga@gt3.bme.hu



DR. ATTILA PIROS was born in Szolnok, Hungary in 1971. He took his MSc degree in the Budapest University of Technology and Economics (BME) in 1995. He founded a small engineering company then took his PhD degree. Now

he works for the BME and he is the responsible for education of CAD design. He has researchers in application of fuzzy method in mechanical engineering, analysis and simulation of the design processes and 3-dimensional reconstruction of the surfaces. His email address is: attila.piros@gt3.bme.hu and his Web-page is: <http://gt3.bme.hu/apiros>



MR. BALÁZS VIDOVICS studied Industrial Design Engineering at the Budapest University of Technology and Economics (BME) and obtained his masters degree in 2003. Ever since he is a lecturer and assistant researcher at the

Department of Machine and Product Design at BME. He is currently a doctoral candidate at the University of West Hungary. His research focus is creativity and innovation in the design process, design thinking and the methodology support in the early phases of the design process. He is also a partner in a design consultancy. His e-mail address is: vidovics.balazs@gt3.bme.hu and his Web-page can be found at <http://gt3.bme.hu/bvidovics>

Electrical and Electromechanical Engineering

DETERMINATION OF AN OPTIMAL SHAPE OF ROTOR FOR THE SYNCHRONOUS REACTIVE FREQUENCY DOUBLER

Aleksandrs Mesņajevs
Elena Ketnere
Faculty of Power and Electrical Engineering
Riga Technical University
1 Blvd Kronvalda, 1010 Riga, Latvia
E-mail: kbl@inbox.lv., ketnere@eef.rtu.lv

KEYWORDS

Finite element method, QuickField, frequency converter, frequency doubler.

ABSTRACT

While developing an electrical machine, the main task is to obtain an optimal magnetic field's distribution. It can be reduced to the selection of such constructive parameters that ensure the best circumstances for magnetic field's existence.

This work is dedicated to the determination of an optimal rotors shape for the synchronous reactive frequency doubler (SRFD). The task is solved by using the finite element method implemented in QuickField software. As a result the rotor's optimal shape is obtained which ensures the highest increased frequency EMF induction.

While studying electrical machine's magnetic field, it is necessary to face analysis tasks, as well as synthesis tasks.

Analysis includes clarification of how different parameters (for example: current, magnetic system's separate parts shape and geometrical dimensions, characteristics of different magnetic materials etc.) influence the magnetic field's character, the field-dependent parameters. Synthesis implies receiving the necessary magnetic field's distribution, which ensures defined/assigned characteristics (for example: magnetic flux, electromotive force, electromagnetic torque, etc.)

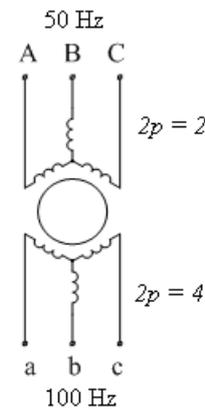
Along with the other multiple questions, specific requirements are being raised with regard to induction distribution in the air gap and the higher harmonics maintenance in fixed constrains.

INTRODUCTION

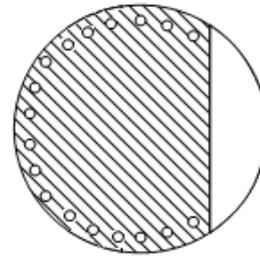
The SRFD is the synchronous reactive machine, which uses the higher harmonics of the magnetic field (the second harmonic). In slots of the SRFD, two windings are placed: the primary, which is connected to the industrial frequency AC network, and the secondary, which is used to receive the increased frequency. It is possible to note, that the synchronous reactive frequency converter is the one-machine aggregate, in which the synchronous reactive motor (stator's primary winding – salient pole rotor) and the inductor generator (salient pole rotor – secondary winding) are combined together.

The primary winding is consuming magnetizing current, which produces the rotating magnetic field in the air gap. From induction's distribution curve, the necessary (second) harmonic is used due to the specific form of the rotor's magnetic system and due to the appropriately selected width of the air gap.

This harmonic induces the increased frequency EMF in the secondary winding. To achieve this, the secondary winding's step must be equal (or almost equal) to the higher harmonic pole pitch of the necessary field.



Figures. 1. Stator winding's scheme

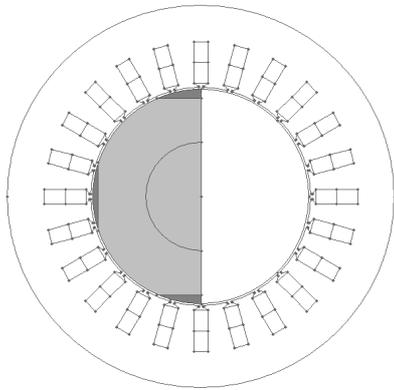


Figures. .2. Rotor's cross-section view

Figure 2 schematically represents the rotor's cross-sectional view. The rotor's parts which are made of the ferromagnetic material are marked with stripes; the unmarked parts are made of the nonmagnetic material (aluminum, plastic). To start-up the SRFD in the rotor's magnetic part, the starting winding ("squirrel cage" shaped, made of aluminum bars and short-circuit rings) is placed.

Power with help of magnetic field is transferred from the primary winding to the secondary winding by means of specific transformation. In this case, the link between the primary and the secondary winding is not provided by the mutual induction flux, but by a part of it – the higher harmonic exuded flux. Power transfer is depending on the geometrical shape of the converter. By having one stator (core, stator windings, slot number, etc.), but different rotors, a very different EMF value can be obtained. So it is necessary to determine an optimal shape of the rotor to receive the highest secondary winding's EMF induction.

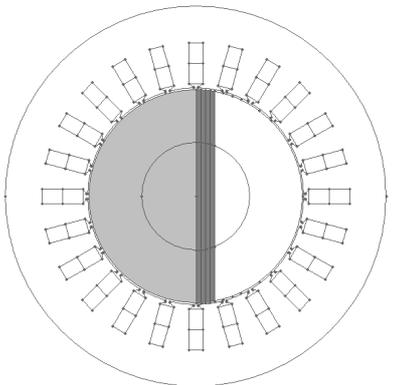
To solve this task, two different approaches were selected. First, it is necessary to clarify the influence of the air gap's form on the second harmonic induction in the air gap.



Figures. 3. SRFD air gap's possible variants.

With gray color the fixed ferromagnetic part of a rotor is highlighted; dark gray – variable part (that in some cases is substituted with air)

Second – creation of specific form rotor with optimal ferromagnetic parts filling angle β that ensures such second harmonics distribution, which guarantee maximal EMF E_2 value.



Figures. 4. SRFD ferromagnetic parts filling angle $\beta = 180^\circ, 185^\circ, 190^\circ, 195^\circ, 200^\circ$

The fixed ferromagnetic part of the rotor is colored in grey; the variable part – in dark gray (that, in some cases, is substituted with air)

All calculations are made for the idle running, when the secondary winding's current $I_2 = 0$ A.

THE SELECTED MACHINE'S PARAMETERS

The finite element method (FEM) is chosen as a magnetic field's research method which is realized in QuickField software. The FEM allows taking into account easily different factors such as configuration of a magnetic system and ferromagnetic non-linear magnetization curves.

For the research, the magnetic system of a commercial asynchronous machine is chosen, in which slots the primary and the secondary windings are placed. The same machine is considered in the previous work (Mesņajevs et al. 2013).

Frequency doublers' stator geometrical parameters are presented in the Table 1,

Table 1. Stator geometrical parameters

D , mm	D_a , mm	t_z , mm	b_r , mm	h_r , mm	δ , mm
120	210,5	15,7	13,1	22,6	1

where D_a – stator's outer diameter;

D – stator inner diameter;

t_z – tooth pitch;

b_r – slots width;

h_r – slots height;

δ – air gap between stator and rotor .

For the SRFD, it is important to exude the magnetic field's second harmonic so that it has the biggest value. To achieve this, the rotor's magnetic system's shape and air gap must be chosen to exude the necessary harmonic in the magnetic field, while the other higher harmonics, whenever possible, are equalized to zero.

SELECTION OF THE AIR GAP'S FORM

To determine possibility of increased frequency EMF amplification, by means of changing an air gap's shape, different variants were selected (see fig. 3.):

- reduction of rotor's ferromagnetic part width along direct-axis by 3 mm and without reduction;
- reduction of rotor's ferromagnetic part width along quadrature-axis by 3 mm and without reduction;
- reduction of rotor's ferromagnetic part width along both axes by 3 mm.

As a basis for mathematical simulation, the SRFD with $\beta = 180^\circ$ is selected due to results of the previous work: "In order to receive the highest in the secondary winding, the induced EMF values frequency doublers'

rotor must be made of at least 50% from ferromagnetic material, the rotor's ferromagnetic parts' filling angle β must be between 180° and 200° and the armature current must be as high as possible (in feasible constraint)" (Mesņajevs et al. 2013).

SELECTION OF THE FERROMAGNETIC PARTS' FILLING ANGLE

As already mentioned, in order to receive the highest in the secondary winding, the induced EMF value rotor's ferromagnetic parts' filling angle β must be between 180° and 200° [1].

In this research, the author is focused towards specifying an optimal ferromagnetic parts' filling angle β . To achieve satisfactory results, the step equal to 1° is chosen. In other words, simulations are made for different rotors that differ by $\beta = 1^\circ$ (see ig. 4.)

In both tasks, the secondary winding induced EMF is calculated using the following formula:

$$E = 4,44 f k_w \frac{2p \cdot q \cdot w_{sp}}{a} 2A_m l, \quad (1)$$

where f – frequency;

k_w – winding coefficient;

q – slots number per pole and phase;

w_{sp} – coil's turn number;

a – parallel turn number;

A_m –specific harmonic's vector

magnetic potential amplitude value

(determined using magnetic field mathematical simulation);

l – machine's length in axial direction.

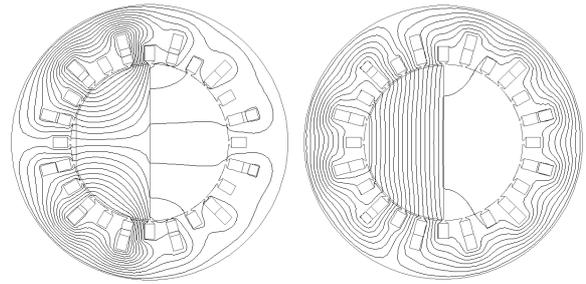
Solution of mentioned tasks is carried out using finite element method for magnetic field investigation, developing and improving software for SRFC's characteristics determination.

SRFD'S SIMULATION RESULTS

Selection of air gap's form

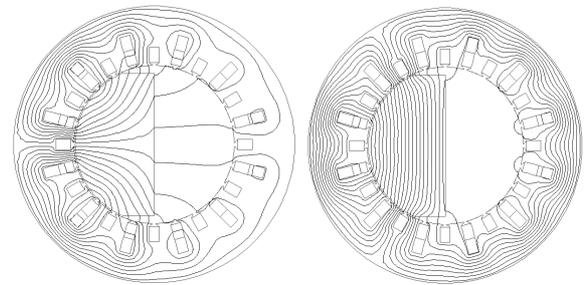
As a result of simulation, the magnetic field pictures are obtained. Five different variants are presented below:

- I. the rotor's ferromagnetic part width along direct-axis is decreased by 3 mm;



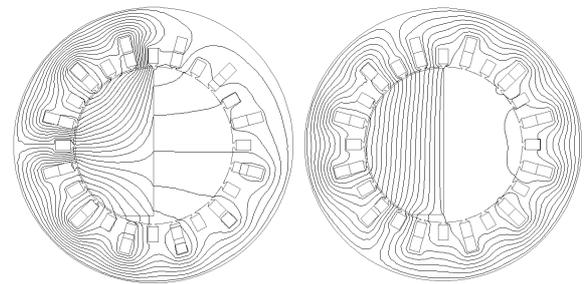
Figures. 5. SRFD magnetic field's pictures for decreased rotor along direct-axis by 3 mm
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

- II. the rotor's ferromagnetic part width along quadrature-axis is decreased by 3 mm on each side;



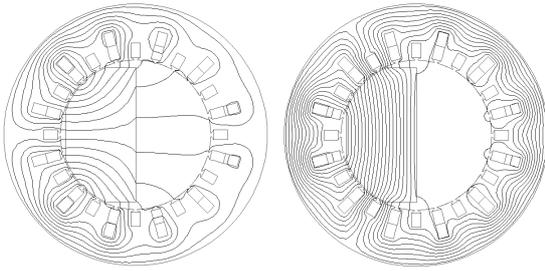
Figures. 6. SRFD magnetic field's pictures for the decreased rotor along quadrature-axis by 3 mm on each side
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

- III. the rotor's ferromagnetic part width along quadrature -axis is decreased by 3 mm only on one side;



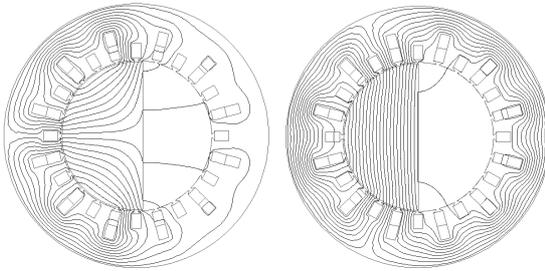
Figures. 7. SRFD magnetic field's pictures for the decreased rotor along quadrature-axis by 3 mm only on side
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

IV. the rotor's ferromagnetic part width is decreased by 3 mm along both axes;



Figures. 8. SRFD magnetic field's pictures for the decreased rotor along both axes by 3 mm
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

V. rotor without reduction.



Figures. 9. The SRFD magnetic field's pictures for the rotor without reduction
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

Simulation is made when the armature current is set $I_a = 72$ A and results are presented in table 2.

Table 2. The secondary winding's EMFs for different rotor shapes.

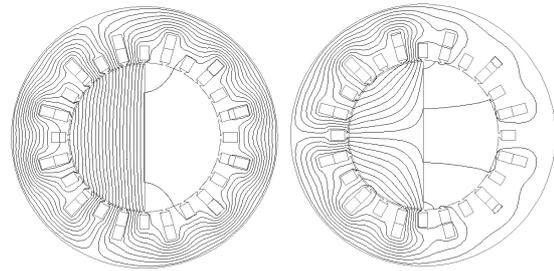
	Condition's number				
	I	II	III	IV	V
E_{2q} , V	6.98	6.72	7.52	4.82	8.99
E_{2d} , V	9.46	10.19	10.15	10.16	9.48
E_2 , V	11.76	12.20	12.63	11.25	13.07

As it is seen in table 2, a reduction of the rotor's ferromagnetic material leads to a decrease of the increased frequency EMF E_2 .

Selection of the rotor's ferromagnetic parts' filling angle

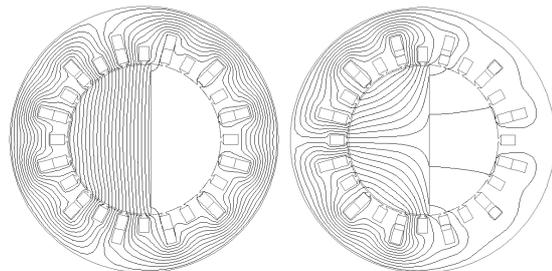
To conduct magnetic field's mathematical modeling, the following magnetic field's pictures are received for different rotor's filling angles β .

1. The rotor's ferromagnetic parts filling angle $\beta = 180^\circ$;



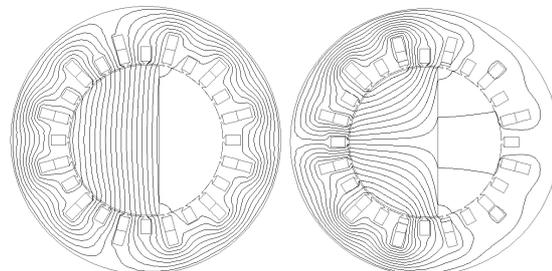
Figures. 10. SRFD magnetic field's pictures for $\beta = 180^\circ$
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

2. The rotor's ferromagnetic parts filling angle $\beta = 190^\circ$



Figures. 11. SRFD magnetic field's pictures for $\beta = 190^\circ$
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

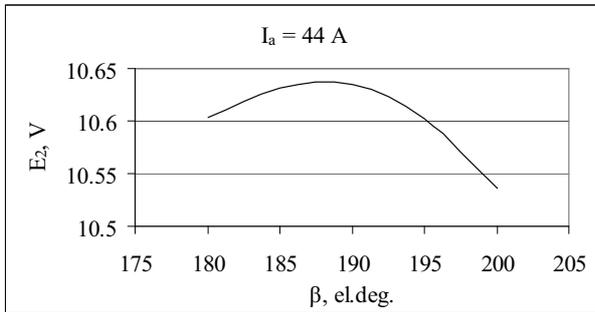
3. The rotor's ferromagnetic parts filling angle $\beta = 200^\circ$



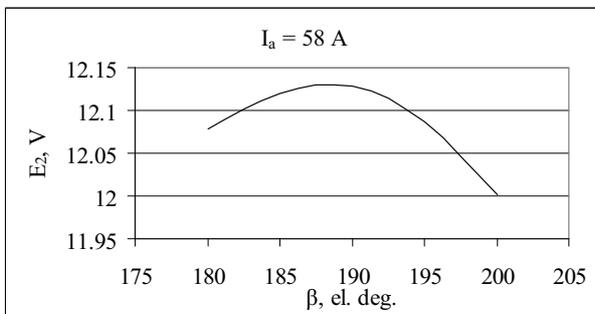
Figures. 12. SRFD magnetic field's pictures for $\beta = 200^\circ$
Left figure- direct-axis ($I_a = I_d$ and $I_q = 0$), right figure – quadrature-axis ($I_a = I_q$ and $I_d = 0$)

Calculations are made for different armature current values $I_a = 44; 58; 72$ A.

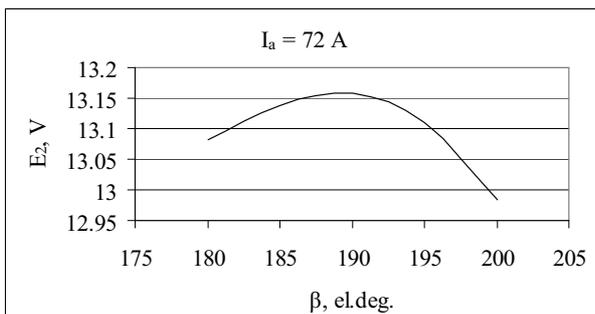
Results of calculations for different currents are presented in fig. 13-15.



Figures. 13. In secondary winding induced EMF's dependence form rotor's ferromagnetic parts filling angle β , when $I_a = 44$ A.



Figures. 14. In secondary winding EMF's dependence form rotor's ferromagnetic parts filling angle β , when $I_a = 58$ A.



Figures. 15. In secondary winding induced EMF's dependence form rotor's ferromagnetic parts filling angle β , when $I_a = 72$ A.

As a result, it can be stated that an optimal rotor's ferromagnetic parts' filling angle $\beta = 187^\circ$.

CONCLUSIONS

- 1) A reduction of rotor's ferromagnetic material leads to a decrease of the increased frequency EMF E_2 ;
- 2) An optimal rotor's ferromagnetic parts filling angle that ensures maximal EMF E_2 value is equal to 187° .

REFERENCES

- Mesņajevs, A. Zviedris, A. Podgornovs. 2011. "Determination of Synchronous Machine's Characteristics Based on the Results of the Mathematical Modelling of the Magnetic Field". *25th European Conference on Modelling and Simulation*, (Jun), 175-180.
- Mesņajevs, A. Zviedris, A. Ketnere, E. 2013. "Selection of synchronous reactive frequency converter's secondary windings parameters and optimization of rotors geometrical dimensions to ensure highest increased frequency EMF induction.". *27th European Conference on Modelling and Simulation*, (May), 764-768
- Zviedris A. 1984. *Elektriskās mašīnas*. Rīga. Zvaigzne
- Таращанский М. 1962. *Синхронно-реактивные преобразователи частоты*. Киев. Государственное издательство технической литературы.

ACKNOWLEDGEMENTS

This paper was supported by the European Social Fund project "Assessment of wind energy potential in Latvia and environmental impact from wind energy installations", No. 2014/0010/1DP/1.1.1.2.0/13/APIA/VIAA/033.

AUTHOR BIOGRAPHIES



ALEKSANDRS MESŅAJEVS is born in 1985, in Rīga, Latvia. In 2012, he graduated Riga Technical University, obtaining the Dr.Sc.ing. degree..

Since 2006 he is working as a laboratory assistant and a scientific assistant; since 2012 – as a lecturer in the Department of Electrical Machines and Apparatus of Riga Technical University.



ELENA KETNERE, Asoc.Prof., Dr.Sc.Eng.

Riga Technical University, Faculty of Power and Electrical Engineering, Department of Electrical Machines and Apparatus; associate professor since 2006), doctoral degree since 2002).

Publications: Simulation of Gas-Turbine Driven Device. The Research of Stability of Synchronization Process with Unitrol 1000 Application. The Research of Stability of Synchronization Process with Mathematical Model's Application of Synchronous Generators.

MODELLING OF BROADBAND ELECTRIC FIELD PROPAGATION IN NONLINEAR DIELECTRIC MEDIA

Matteo Conforti
PhLAM/IRCICA UMR 8523/USR 3380
CNRS-Université Lille 1
F-59655 Villeneuve d'Ascq, France
Email: conforti.matteo@gmail.com

Fabio Baronio
Dipartimento di Ingegneria dell'Informazione
Universtità di Brescia
Via Branze 38, 25123 Brescia, Italy
Email: fabio.baronio@unibs.it

KEYWORDS

Ultrafast nonlinear optics; frequency conversion

ABSTRACT

We derive unidirectional pulse propagation equations to describe extreme high-intensity and ultra-broadband optical interactions in uniaxial crystals, showing both second- and third-order nonlinear optical susceptivities. Two nonlinearly coupled first order (in the propagation coordinate) equations describe the dynamics and interactions of the ordinary and extraordinary field polarizations, and are valid for arbitrarily wide pulse bandwidth. We exploit this model to predict harmonic and supercontinuum generation in BBO crystals under strong and competing influence of quadratic and cubic susceptivities.

INTRODUCTION

In recent years there has been a great deal of interest in research on second-harmonic (SHG) (Mironov et al. 2009), high-order harmonic (HHG) (Krauzs and Ivanov 2009), and supercontinuum (SC) generation (Dudley et al. 2006) in nonlinear optical media for such diverse applications as frequency metrology, few-cycle pulse generation, spectroscopy, biological and medical analyses.

The SHG of super-strong ultrashort (tens of femtoseconds) laser pulses, using the $\chi^{(2)}$ nonlinearities in optical crystals, is a very important task, because the process can be used not only for wavelength conversion, but for significant improvement of temporal intensity contrast ratio and pulse shortening. SHG is especially important for Ti:sapphire laser facilities operating at 800 nm (Aoyama et al. 2003) and optical parametric amplifiers at 910 nm (Lozhkarev et al. 2007).

SC generation has been performed conventionally using the $\chi^{(3)}$ nonlinearities in optical fibers. Due to the high nonlinearity and engineerable dispersion available in fibers, spectra spanning multiple octaves can be achieved (Farrell et al. 2012; Fang et al. 2012). However, reaching the mid-infrared spectral region with $\chi^{(3)}$ -based SC sources is challenging (Price et al. 2007). A promising alternative approach consists

on the exploitation of the $\chi^{(2)}$ nonlinearities of optical crystals for SC generation (Conforti et al. 2010a; Phillips et al. 2011a). SC interactions can readily be achieved in birefringent or quasi-phase matched (QPM) crystals (Conforti et al. 2010b; Phillips et al. 2011b), with high-intensity light pulse excitation. Quadratic SC generation, difference frequency generation, optical parametric generation and QPM engineering are currently active areas of research (Conforti et al. 2007a; Zhou et al. 2012; Levenius et al. 2012).

Nowadays, technological advances in ultrafast optics have permitted the generation of ultraintense light pulses comprising merely a few field oscillation cycles. Peak intensities approaches 10^{15}W/cm^2 (Sung et al. 2010), opening the study of an entirely new realm of nonlinear interactions in solid materials.

Beta-Barium-Borate ($\beta\text{-BaB}_2\text{O}_4$, BBO) is a very popular crystal, among all solid-state optical materials: BBO has a high damage threshold, low dispersion and $\chi^{(2)}$ nonlinearities of few pm/V allowing for efficient quadratic frequency conversion interactions (Nikogosyan 2005).

In this work, we explore the use of BBO crystals in extreme optical regimes, where dispersion effects and cubic nonlinearities play an essential role. In particular, we derive a comprehensive model to describe the propagation of extreme high-intensity and ultra-broadband optical pulses in BBO crystals. This model provides a powerful tool due to its generality and simplicity, and can be easily solved with a modest computational effort.

The paper is organized as follows. In Section 2, we recall the derivation of the master equations in uniaxial media, discussing the validity of the model. We consider both the second- and third-order nonlinear contributions, and their angular dependences. We take into account all possible second- and third-order interactions, including ones typically non-phase-matchable. In Section 3, we present some numerical examples of second harmonic generation and supercontinuum generation in BBO crystals, showing the key role of cubic susceptibility. Eventually we draw our conclusions in Section 4.

DERIVATION OF THE MODEL

In this section we review and extend the derivation of the unidirectional nonlinear vector field equations reported in (Conforti et al. 2011) (also called Forward Maxwell Equations, FME (Housakou and Herrmann 2001), or Unidirectional Pulse Propagation Equation, UPPE (Kolesik et al. 2002)), describing the propagation of the ordinary and extraordinary polarizations of the electric field in uniaxial crystals with both $\chi^{(2)}$ and $\chi^{(3)}$ nonlinearities.

We start from Maxwell equations written in MKS units, in the reference frame $x'y'z'$

$$\nabla' \times \mathbf{E}' = -\frac{\partial \mathbf{B}'}{\partial t} \quad (1)$$

$$\nabla' \times \mathbf{H}' = \frac{\partial \mathbf{D}'}{\partial t} \quad (2)$$

$$\mathbf{B}' = \mu_0 \mathbf{H}' \quad (3)$$

$$\mathbf{D}' = \mathbf{D}'_L + \mathbf{P}'_{NL} \quad (4)$$

where \mathbf{D}'_L and \mathbf{P}'_{NL} account for the linear and nonlinear response of the medium, respectively. The components of the linear displacement vector for a dispersive anisotropic medium reads (assuming summation over repeated indexes)

$$D'_{L,j} = \varepsilon_0 \int_{-\infty}^{\infty} \varepsilon'_{jk}(t-t') E'_k(t') dt' \quad (5)$$

In the reference frame of the principal axes of a uniaxial crystal, the dielectric permittivity tensor is the diagonal matrix $\varepsilon = \text{diag}(\varepsilon_o, \varepsilon_o, \varepsilon_e)$, where $\varepsilon_o, \varepsilon_e$ are the ordinary and extraordinary relative dielectric permittivity, respectively. The reference frame of the principal axes of the crystal ($x'y'z'$) is not convenient for the derivation of the propagation equations. We introduce a reference frame xyz that is rotated by (θ, ϕ) with respect to crystal axes. Namely, θ is the angle between the propagation vector (parallel to z) and the crystalline z' axis (the crystal optical axis), and ϕ is the azimuthal angle between the propagation vector and the $x'z'$ crystalline plane. The two reference frames are linked by the orthogonal rotation matrix A :

$$A = \begin{bmatrix} \cos \phi \cos \theta & \sin \phi \cos \theta & -\sin \theta \\ -\sin \phi & \cos \phi & 0 \\ \sin \theta \cos \phi & \sin \theta \sin \phi & \cos \theta \end{bmatrix}. \quad (6)$$

The dielectric permittivity tensor in the xyz frame is no longer diagonal, and it can be written as

$$\begin{aligned} \varepsilon &= A \varepsilon' A^T \\ &= \begin{bmatrix} \varepsilon_o \cos^2 \theta + \varepsilon_e \sin^2 \theta & 0 & (\varepsilon_o - \varepsilon_e) \cos \theta \sin \theta \\ 0 & \varepsilon_o & 0 \\ (\varepsilon_o - \varepsilon_e) \cos \theta \sin \theta & 0 & \varepsilon_o \sin^2 \theta + \varepsilon_e \cos^2 \theta \end{bmatrix}. \end{aligned} \quad (7)$$

In the reference frame xyz , it is possible to decompose the electromagnetic field into two linear and orthogonal polarizations of \mathbf{D} , both transverse to the propagation direction z (Landau and Lifshitz 1984): $\mathbf{D} = (0, D_y, 0)^T + (D_x, 0, 0)^T$. We assume the propagation of plane waves, so the electric field and displacement vectors depend upon the z coordinate

(and time) only. It is worth noting that this decomposition is rigorous for linear propagation only, since the nonlinearity can rotate locally the polarization. However it is reasonable to consider the nonlinearity as a perturbative term whose effect is to couple the orthogonal polarized field vector components during propagation. If we neglect dispersion and nonlinearity, just for the moment, the electric field vector can be straightforwardly computed as:

$$\mathbf{E} = \varepsilon_0^{-1} \varepsilon^{-1} \mathbf{D} = \varepsilon_0^{-1} \begin{bmatrix} \left(\frac{\cos^2 \theta}{\varepsilon_o} + \frac{\sin^2 \theta}{\varepsilon_e} \right) D_x \\ \varepsilon_o^{-1} D_y \\ \frac{\varepsilon_e - \varepsilon_o}{\varepsilon_e \varepsilon_o} \cos \theta \sin \theta D_x \end{bmatrix} \quad (8)$$

By eliminating the magnetic field from Maxwell equations we obtain the vector wave equation:

$$\nabla \times \nabla \times \mathbf{E} - \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 \mathbf{D}_L}{\partial t^2} = \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 \mathbf{P}_{NL}}{\partial t^2} \quad (9)$$

Note that obviously $\nabla \cdot \mathbf{D} = 0$, but $\nabla \cdot \mathbf{E} \neq 0$. By writing (9) in components we obtain

$$\frac{\partial^2 E_x}{\partial z^2} - \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 D_{L,x}}{\partial t^2} = \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 P_{NL,x}}{\partial t^2} \quad (10)$$

$$\frac{\partial^2 E_y}{\partial z^2} - \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 D_{L,y}}{\partial t^2} = \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 P_{NL,y}}{\partial t^2} \quad (11)$$

$$0 = \frac{1}{\varepsilon_0 c^2} \frac{\partial^2 P_{NL,z}}{\partial t^2} \quad (12)$$

The last equation witnesses the fact that the decomposition into two independent orthogonal polarizations is rigorous only in the linear case. We neglect $P_{NL,z}$, in the reasonable hypothesis of small nonlinearity.

Exploiting the relation (5) we obtain:

$$\begin{aligned} \frac{\partial^2 E_m(z, t)}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \int_{-\infty}^{+\infty} E_m(z, t') \varepsilon_m(t-t') dt' \\ = \frac{1}{\varepsilon_0 c^2} \frac{\partial^2}{\partial t^2} P_{NL,m}(z, t), \quad m = x, y \end{aligned} \quad (13)$$

where we have defined

$$\varepsilon_x = \left(\frac{\cos^2 \theta}{\varepsilon_o} + \frac{\sin^2 \theta}{\varepsilon_e} \right)^{-1} \quad (14)$$

$$\varepsilon_y = \varepsilon_o \quad (15)$$

We thus have obtained the propagation equations for an ordinary polarized wave E_y and an extraordinary polarized wave E_x .

By defining the Fourier transform $\mathcal{F}[E](\omega) = \hat{E}(\omega) = \int_{-\infty}^{+\infty} E(t) e^{-i\omega t} dt$, we can write (13) in the frequency domain:

$$\frac{\partial^2 \hat{E}_m(z, \omega)}{\partial z^2} + \frac{\omega^2}{c^2} \hat{\varepsilon}_m(\omega) \hat{E}_m(z, \omega) = -\frac{\omega^2}{\varepsilon_0 c^2} \hat{P}_{NL,m}(z, \omega), \quad (16)$$

where c is the velocity of light in vacuum, ε_o is the vacuum dielectric permittivity, $\hat{\varepsilon}_m(\omega) = 1 + \hat{\chi}_m(\omega)$, $\hat{\chi}_m(\omega)$ is the linear electric susceptibility and $k_m(\omega) = (\omega/c) \sqrt{\hat{\varepsilon}_m(\omega)}$ is the propagation wavenumber.

We now proceed to obtain, from the second order vector wave equation (16), an equation, first order in the propagation coordinate z , describing electromagnetic fields propagating in the forward direction only. Several techniques have been proposed in literature in order to achieve a pulse propagation equation with minimal assumptions (Brabec and Krausz 2000; Housakou and Herrmann 2001; Kolesik et al. 2002; Kolesik and Moloney 2004; Genty et al. 2007; Kinsler et al. 2005; Kinsler 2010; Kumar 2010). The interested reader can find in (Kinsler 2010; Kolesik et al. 2012) an exhaustive discussion on the different derivation styles. Here we decided to follow the approach outlined in the review paper (Kolesik et al. 2012), that combines minimal assumptions and straightforward derivation.

We write the electric field components in spectral domain as the sum of a forward (F) and a backward (B) propagating part, that with our definition of the Fourier transform reads:

$$\hat{E}_m(z, \omega) = \hat{F}_m(z, \omega)e^{-ik_m(\omega)z} + \hat{B}_m(z, \omega)e^{ik_m(\omega)z}. \quad (17)$$

By plugging Ansatz (17) into (16), we get:

$$\left(\frac{\partial^2 \hat{F}_m}{\partial z^2} - 2ik_m(\omega) \frac{\partial \hat{F}_m}{\partial z} \right) e^{-ik_m(\omega)z} + \left(\frac{\partial^2 \hat{B}_m}{\partial z^2} + 2ik_m(\omega) \frac{\partial \hat{B}_m}{\partial z} \right) e^{ik_m(\omega)z} = -\frac{\omega^2}{\varepsilon_0 c^2} \hat{P}_{NL,m},$$

that can be rewritten as:

$$\begin{aligned} & \frac{\partial}{\partial z} \left(\frac{\partial \hat{F}_m}{\partial z} e^{-ik_m(\omega)z} + \frac{\partial \hat{B}_m}{\partial z} e^{ik_m(\omega)z} \right) - \\ & - ik_m(\omega) \left(\frac{\partial \hat{F}_m}{\partial z} e^{-ik_m(\omega)z} - \frac{\partial \hat{B}_m}{\partial z} e^{ik_m(\omega)z} \right) = \\ & -\frac{\omega^2}{\varepsilon_0 c^2} \hat{P}_{NL,m}, \end{aligned} \quad (18)$$

from where it is trivial to see that vector wave equation (16) is satisfied exactly, if the forward and backward components satisfy the following first order equations:

$$\begin{aligned} \frac{\partial \hat{F}_m(z, \omega)}{\partial z} &= -\frac{i}{2k_m(\omega)} \frac{\omega^2}{\varepsilon_0 c^2} \hat{P}_{NL,m}(z, \omega) e^{+ik_m(\omega)z} \\ \frac{\partial \hat{B}_m(z, \omega)}{\partial z} &= +\frac{i}{2k_m(\omega)} \frac{\omega^2}{\varepsilon_0 c^2} \hat{P}_{NL,m}(z, \omega) e^{-ik_m(\omega)z}. \end{aligned} \quad (19)$$

It is worth noting that up to this point we did not make any assumptions, so the model is equivalent to the starting equations. Equations (19) represent a nonlinear boundary value problem that cannot be solved with direct methods, but must be solved iteratively. However in the great majority of cases of interest, we can assume that (i) there are no reflections and (ii) that nonlinear polarization does not couple forward and backward waves (perturbative regime). In this case we can assume $\hat{B}_m(z, \omega) \approx 0$ and Eqs. (19), through (17), reduce to the Forward Maxwell Equations:

TABLE I
EFFECTIVE QUADRATIC NONLINEAR COEFFICIENTS. $d_{22} = 2.2\text{PM/V}$,
 $d_{31} = 0.04\text{PM/V}$.

Coefficient	Expression	Interaction
d_0	$-3d_{31} \cos^2 \theta \sin \theta - d_{22} \cos^3 \theta \sin 3\phi$	eee
d_1	$-d_{22} \cos 3\phi \cos^2 \theta$	eeo, oeo, oee
d_2	$-d_{31} \sin \theta + d_{22} \cos \theta \sin 3\phi$	ooe, eoo, oeo
d_3	$d_{22} \cos 3\phi$	ooo

$$\frac{\partial \hat{E}_m(z, \omega)}{\partial z} + ik_m(\omega) \hat{E}_m(z, \omega) = -i \frac{\omega}{2\varepsilon_0 c n_m(\omega)} \hat{P}_{NL,m}(z, \omega). \quad (20)$$

We consider an instantaneous nonlinear polarization composed of a quadratic and cubic parts (summation over repeated indexes is assumed)

$$P'_{NL,j} = \varepsilon_0 (\chi_{jkl}^{(2)} E'_k E'_l + \chi_{jklm}^{(3)} E'_k E'_l E'_m), \quad (21)$$

where $\chi_{jkl}^{(2)}$ and $\chi_{jklm}^{(3)}$ are the second and third order nonlinear susceptibility tensors, that are usually given in the crystal axes reference frame. In order to obtain the effective nonlinearity (Midwinter and Warner 1965a; Midwinter and Warner 1965b), we have to rotate the polarization vector with matrix A , following the prescription

$$\mathbf{P}_{NL}(\mathbf{E}) = A \mathbf{P}'_{NL}(A^T \mathbf{E}). \quad (22)$$

After some calculations, we can write:

$$\frac{\partial \hat{E}_x}{\partial z} + ik_x(\omega) \hat{E}_x = \frac{-i\omega}{c n_x(\omega)} \hat{P}_x \quad (23)$$

$$\frac{\partial \hat{E}_y}{\partial z} + ik_y(\omega) \hat{E}_y = \frac{-i\omega}{c n_y(\omega)} \hat{P}_y$$

where the nonlinear terms P_x, P_y read as follows:

$$\begin{aligned} P_x &= d_0 E_x^2 + 2d_1 E_x E_y + d_2 E_y^2 \\ &+ \frac{1}{2} (c_0 E_x^3 + 3c_1 E_x^2 E_y + 3c_2 E_y^2 E_x + c_3 E_y^3), \end{aligned} \quad (24)$$

$$\begin{aligned} P_y &= d_1 E_x^2 + 2d_2 E_x E_y + d_3 E_y^2 \\ &+ \frac{1}{2} (c_1 E_x^3 + 3c_2 E_x^2 E_y + 3c_3 E_y^2 E_x + c_4 E_y^3). \end{aligned} \quad (25)$$

where $d_m, m = 0, \dots, 3$, are the effective nonlinearity for quadratic interactions, whereas $c_m, m = 0, \dots, 4$ are the effective cubic nonlinearities. The values of the effective nonlinearity depend upon the crystal and their values can be found in literature (Nikogosyan 2005; Midwinter and Warner 1965a; Midwinter and Warner 1965b; Banks et al. 2002). In Tables I, II we report the effective nonlinearity for the crystals of class $3m$, to which BBO belongs, and specify the kind of interaction. For example, eeo ($e + e \rightarrow o$) indicates the sum frequency generation of two extraordinarily polarized electric fields (E_x) that generate an ordinarily polarized field (E_y).

Equations (23) are first order in the propagation coordinate, conserve the total field energy and retain their validity for

TABLE II

EFFECTIVE CUBIC NONLINEAR COEFFICIENTS. $c_{11} = 5.6 \cdot 10^{-22} \text{M}^2/\text{V}^2$, $c_{10} = -0.24 \cdot 10^{-22} \text{M}^2/\text{V}^2$, $c_{16} = -1.4 \cdot 10^{-22} \text{M}^2/\text{V}^2$ (BACHE ET AL. 2013).

Coefficient	Expression	Interaction
c_0	$c_{11} \cos^4 \theta + c_{33} \sin^4 \theta + \frac{3}{2} c_{16} \sin^2 2\theta$ $-4c_{10} \sin 3\phi \sin \theta \cos^3 \theta$	eeee
c_1	$\frac{3}{2} c_{10} \cos 3\phi \sin 2\theta \cos \theta$	eeoe, eeco
c_2	$-\frac{1}{3} c_{11} \cos^2 \theta + c_{16} \sin^2 \theta$ $+c_{10} \sin 2\theta \sin 3\phi$	ooee, eooo
c_3	$c_{10} \cos 3\phi \sin \theta$	oooo, ooeo
c_4	c_{11}	oooo

arbitrary wide pulse bandwidth. The computational effort needed to solve these equations, by a standard split step Fourier method exploiting Runge-Kutta for the nonlinear step, is of the order of magnitude of that needed for solving the standard three-wave equations universally exploited to describe light propagation in quadratic crystals (Conforti et al. 2007b; Baronio et al. 2008; Baronio et al. 2010). However Eqs. (23) are far more general, and are equivalent to Maxwell equations when dealing with unidirectional propagation (Kolesik et al. 2002; Kolesik et al. 2012).

EXAMPLES

In this section, we first show a representative example of the modeling of supercontinuum generation by means of competing quadratic and cubic nonlinearities. Then, we present simulations of soliton compression and blue-shifted dispersive waves generation in BBO.

Supercontinuum generation

We fix the orientation angles of the BBO crystal to $\theta = 19^\circ$ and $\phi = 90^\circ$.

We consider the propagation of an ordinarily polarized pulse of duration $T = 20$ fs, peak intensity of $120 \text{GW}/\text{cm}^2$, central wavelength $\lambda_0 = 1200$ nm, where BBO shows normal dispersion ($\beta'' = 0.27 \text{ps}^2/\text{m}$). Under such assumptions, considering a type I ($o + o \rightarrow e$) quadratic interaction, the mismatch is $\Delta k = k_e(2\omega) - 2k_o(\omega) = 3.3 \cdot 10^4 \text{m}^{-1}$, that give rise to an effective cascaded negative (defocusing) Kerr nonlinearity. The combination of normal dispersion and defocusing nonlinearity allows for solitary wave propagation. However, intrinsic cubic nonlinearities in the material are self-focusing and can compete with the induced quadratic self-defocusing effects (Bache et al. 2008).

The cascaded quadratic and cubic Kerr nonlinearities are expressed as $\gamma_2 = -(\frac{\omega d_{eff}}{nc})^2 \frac{1}{\Delta k} [m/V^2]$ and $\gamma_3 = \frac{3}{8} \frac{\omega c_{eff}}{nc} [m/V^2]$, with d_{eff} and c_{eff} effective nonlinear coefficients of Tables I, II. In the present case we find that the strongest interactions are $o + o \rightarrow e$ (quadratic), and $o + o + o \rightarrow o$ (cubic), so we can approximate $d_{eff} \approx d_2$ and $c_{eff} \approx c_4$.

Figure 1a shows the time domain evolution of the ordinarily polarized (o) electric field envelope during the propagation in BBO crystal. With envelope we mean the inverse Fourier transform of the positive frequency components of the spectrum.

This visualization permits to have an envelope-like appearance, without fast oscillations of the carrier, but accounts of all frequency components. The input pulse undergoes a strong compression up to $z = 0.6$ mm, where the minimum pulse duration and maximum of spectral extension is achieved. Figure 1b shows the evolution of the ordinarily polarized field spectrum. The compression is due to the cascaded quadratic effects ($\gamma_2 = -14 \cdot 10^{-16} \text{m}/\text{V}^2$, $\gamma_3 = 6.65 \cdot 10^{-16} \text{m}/\text{V}^2$). At the compression point the ordinary polarized pulse shows trailing oscillations, and subsequently radiation is emitted at a slower group velocity: a linear dispersive wave located in the red part of the spectrum at 2400 nm (Bache et al. 2010a; Bache et al. 2010b).

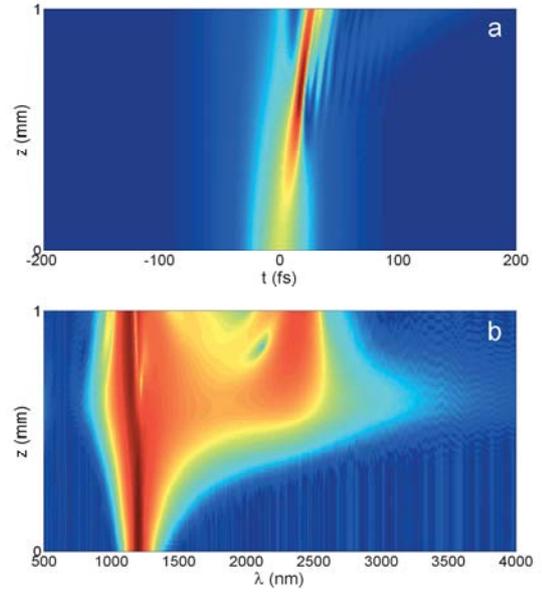


Fig. 1. (Color online) a) temporal propagation and b) field spectrum evolution (decibels) of the ordinarily polarized electric field envelope in BBO crystal. The initial pulse has duration $T = 20$ fs, wavelength $\lambda_0 = 1200$ nm, peak intensity of $120 \text{GW}/\text{cm}^2$. Crystal's orientation: $\theta = 19^\circ$ and $\phi = 90^\circ$.

Then, we decrease the θ orientation angle of the BBO crystal to $\theta = 16.2^\circ$ ($\phi = 90^\circ$), keeping fixed the input pulse characteristics. In this case the dispersion is unaltered ($\beta'' = 0.27 \text{ps}^2/\text{m}$), but the mismatch is $\Delta k = 7.1 \cdot 10^4 \text{m}^{-1}$.

Figure 2a shows the time domain evolution of the ordinarily polarized electric field during the propagation in BBO crystal, whereas figure 2b shows the evolution of the field spectrum. The scenario has been dramatically changed with respect to the previous case. In fact, the effective quadratic negative Kerr nonlinearity ($\gamma_2 = -6.6 \cdot 10^{-16} \text{m}/\text{V}^2$), induced by mismatched type I ($o + o \rightarrow e$) interaction, is perfectly balanced by the cubic nonlinearity of the medium ($o + o + o \rightarrow o$ interaction). The ordinarily polarized pulse propagates in the BBO crystal in the same way as the nonlinearities were vanishing, independently from input intensity.

Soliton compression and emission of resonant radiation

We fix the orientation angles of the BBO crystal to $\theta = 80^\circ$ and $\phi = 90^\circ$.

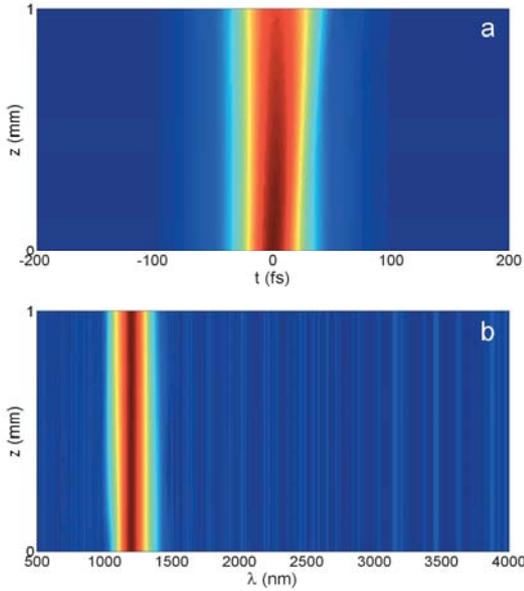


Fig. 2. (Color online) a) temporal propagation and b) field spectrum evolution (decibels) of the ordinarily polarized electric field envelope in BBO crystal. The initial pulse has duration $T = 20$ fs, wavelength $\lambda_0 = 1200$ nm, peak intensity of 120 GW/cm^2 . Crystal's orientation $\theta = 16.2^\circ$ and $\phi = 90^\circ$.

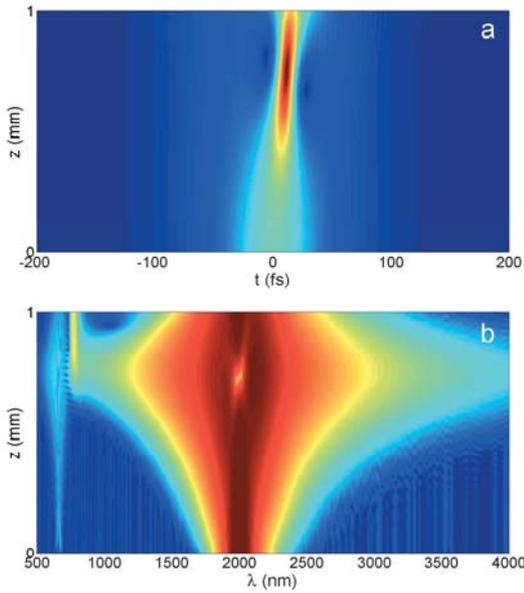


Fig. 3. (Color online) a) temporal propagation and b) field spectrum evolution of the ordinarily polarized electric field envelope in BBO crystal. The initial pulse has duration $T = 30$ fs, wavelength $\lambda_0 = 2000$ nm, peak intensity of 130 GW/cm^2 . Crystal's orientation: $\theta = 80^\circ$ and $\phi = 90^\circ$.

We consider the propagation of an ordinarily polarized pulse of duration $T = 30$ fs, with intensity of 130 GW/cm^2 , central wavelength $\lambda_0 = 2000$ nm, where BBO shows anomalous dispersion ($\beta'' = -0.09 \text{ ps}^2/\text{m}$). Under such assumptions, considering a quadratic type I ($o + o \rightarrow e$) interaction, the mismatch is $\Delta k = -6 \cdot 10^5 \text{ m}^{-1}$, that give rise to an effective cascaded positive (focusing) Kerr nonlinearity. The combina-

tion of anomalous dispersion and focusing nonlinearity can allow for solitary wave dynamics. The cubic nonlinearities in the material are self-focusing too, and are stronger with respect to the induced cascaded quadratic self-focusing effects: in fact we have $\gamma_2 = 1.8 \cdot 10^{-18} \text{ m/V}^2$, $\gamma_3 = 4 \cdot 10^{-16} \text{ m/V}^2$.

Figure 3a shows the time domain propagation of the ordinarily polarized electric field during the propagation in BBO crystal, figure 3b shows the evolution of the field spectrum. The input pulse undergoes a strong compression up to $z = 0.8$ mm, where the minimum pulse duration and maximum of spectral extension is achieved. The compression is due to high-order cubic soliton excitation. A linear dispersive wave (Wai et al. 1987), located in the blue part of the spectrum at 900 nm, has been generated.

CONCLUSIONS

We have derived unidirectional pulse propagation equations to describe extreme high-intensity and ultra-broadband optical interactions in anisotropic crystals showing both quadratic and cubic nonlinear optical susceptibilities, taking BBO as the most relevant example. This model can be used to model harmonic and ultrabroadband generation in BBO crystals under strong and competing influence of quadratic and cubic susceptibilities.

REFERENCES

- Aoyama M.; K. Yamakawa; Y. Akahane; J. Ma; N. Inoue; H. Ueda; and H. Kiriya. 2003. "0.85-PW, 33-fs Ti:sapphire laser," *Optics Letters* 28, 1594-1596.
- Bache M.; O. Bang; W. Krolikowski; J. Moses; and F. Wise. 2008. "Limits to compression with cascaded quadratic soliton compressors," *Optics Express* 16, 3273-3287.
- Bache M.; O. Bang; B. Zhou; J. Moses; and F. Wise. 2010. "Optical Cherenkov radiation in ultrafast cascaded second-harmonic generation," *Physical Review A* 82, 063806.
- Bache M.; O. Bang; B. Zhou; J. Moses; and F. Wise. 2010. "Optical Cherenkov radiation by cascaded nonlinear interaction: an efficient source of few-cycle energetic near- to mid-IR pulses," *Optics Express* 19, 22557-22562.
- Bache M.; H. Guo; B. Zhou; and X. Zheng. 2013. "The anisotropic Kerr nonlinear refractive index of β -BaB₂O₄ nonlinear crystal," *Optical Materials Express* 3, 357-382.
- Banks P.S.; M. D. Feit; and M. D. Perry. 2002. "High intensity third-harmonic generation," *Journal of the Optical Society of America B* 19, 102-118.
- Baronio F.; M. Conforti; A. Degasperis; and S. Wabnitz. 2008. "Three-wave trapped solitons for tunable high-repetition rate pulse train generation," *IEEE Journal of Quantum Electronics* 44, 542-546.
- Baronio F.; M. Conforti; C. De Angelis; A. Degasperis; M. Andreana; V. Couderc; and A. Barthelemy. 2010. "Velocity-locked solitary waves in quadratic media," *Physical Review Letters* 104, 113902.
- Brabec T. and F. Krausz. 2000. "Intense few-cycle laser fields: Frontiers of nonlinear optics," *Reviews of Modern Physics* 72, 545-591.
- Conforti M.; F. Baronio; and C. De Angelis. 2007. "From femtosecond infrared to picosecond visible pulses: temporal shaping with high-efficiency conversion," *Optics Letters* 32, 1779-1781.
- Conforti M.; F. Baronio; A. Degasperis; and S. Wabnitz. 2007. "Parametric frequency conversion of short optical pulses controlled by a CW background," *Optics Express* 15, 12264-12251.
- Conforti M.; F. Baronio; and C. De Angelis. 2010. "Nonlinear envelope equation for broadband optical pulses in quadratic media," *Physical Review A* 81, 053841.
- Conforti M.; F. Baronio; and C. De Angelis. 2010. "Ultra-broadband optical phenomena in quadratic nonlinear media," *IEEE Photonics Journal* 2, 600-610.

- Conforti M.; F. Baronio; and C. De Angelis. 2011. "Modeling of ultrabroadband and single-cycle phenomena in anisotropic quadratic crystals," *Journal of the Optical Society of America B* 28, 1231-1237.
- Dudley J. M.; G. Genty; and S. Coen. 2006. "Supercontinuum generation in photonic crystal fiber," *Rev. Mod. Phys.* 78, 1135-1184.
- Fang X.; M. Hu; L. Huang; L. Chai; N. Dai; J. Li; A. Tashchilina; A. M. Zheltikov; and C. Wang. 2012. "Multiwatt octave-spanning supercontinuum generation in multicore photonic-crystal fiber," *Optics Letters* 37, 2292-2294.
- Farrell C.; K. A. Serrels; T. R. Lundquist; P. Vedagarbha; and D. T. Reid. 2012. "Octave-spanning super-continuum from a silica photonic crystal fiber pumped by a 386 MHz Yb: fiber laser," *Optics Letters* 37, 1778-1780.
- Genty G.; P. Kinsler; B. Kibler; and J.M. Dudley. 2007. "Nonlinear envelope equation modeling of sub-cycle dynamics and harmonic generation in nonlinear waveguides," *Optics Express* 15 5382-5387.
- Housakou A. V. and J. Herrmann. 2001. "Supercontinuum generation of higher-order solitons by fission in photonic crystal fibers," *Physical Review Letters* 87, 203901.
- Kinsler P.; S. B. P. Radnor; and G. H. C. New. 2005. "Theory of directional pulse propagation," *Physical Review A* 72, 063807.
- Kinsler P. 2010. "Optical pulse propagation with minimal approximations," *Physical Review A* 81, 013819.
- Kolesik M.; J. V. Moloney; and M. Mlejnek. 2002. "Unidirectional optical pulse propagation equation," *Physical Review Letters* 89, 283902.
- Kolesik M. and J. V. Moloney. 2004. "Nonlinear optical pulse propagation simulation: From Maxwell's to unidirectional equations," *Physical Review E* 70, 036604.
- Kolesik M.; P. T. Whalen; and J. V. Moloney. 2012. "Theory and simulation of ultrafast intense pulse propagation in extended media," *IEEE Journal of Selected Topics in Quantum Electronics* 18, 494-506.
- Krausz F.; and M. Ivanov. 2009. "Attosecond physics," *Reviews of Modern Physics* 81, 163-234.
- Kumar A. 2010. "Ultrashort pulse propagation in a cubic medium including the Raman effect," *Physical Review A* 81, 013807.
- Landau L. D. and E. M. Lifshitz. 1984. "Electrodynamics of continuous media", Pergamon, New York.
- Levenius M.; M. Conforti; F. Baronio; V. Pasiskevicius; F. Laurell; C. De Angelis; and K. Gallo. 2012. "Multistep quadratic cascading in broadband optical parametric generation," *Optics Letters* 37, 1727-1729.
- Lozhkarev V. V.; G. I. Freidman; V. N. Ginzburg; E. V. Katin; E. A. Khazanov; A. V. Kirsanov; G. A. Luchinin; A. N. Maloshakov; M. A. Martyanov; O. V. Palashov; A. K. Poteomkin; A. M. Sergeev; A. A. Shaykin; and I. V. Yakovlev. 2007. "Compact 0.56 petawatt laser system based on optical parametric chirped pulse amplification in KDP crystals," *Laser Physics Letters* 4, 421-427.
- Midwinter J. E. and J. Warner. 1965. "The effects of phase matching method and of uniaxial crystal symmetry on the polar distribution of second-order non-linear optical polarization", *British Journal of Applied Physics* 16, 1135-1142.
- Midwinter J. E. and J. Warner 1965. "The effects of phase matching method and of crystal symmetry on the polar dependence of third-order non-linear optical polarization", *British Journal of Applied Physics* 16, 1667-1674 (1965).
- Mironov S.; V. Lozhkarev; V. Ginzburg; and E. Khazanov. 2009. "High-efficiency second-harmonic generation of superintense ultrashort laser pulses," *Applied Optics* 48, 2051-2057.
- Nikogosyan D. N.. 2005. *Nonlinear Optical Crystals: A Complete Survey*, Springer.
- Phillips C. R.; C. Langrock; J. S. Pelc; M. M. Fejer; J. Jiang; M. E. Fermann; and I. Hartl. 2011. "Supercontinuum generation in quasi-phase-matched LiNbO₃ waveguide pumped by a Tm-doped fiber laser system," *Optics Letters* 36, 3912-3914.
- Phillips C. R.; C. Langrock; J. S. Pelc; M. M. Fejer; I. Hartl; and M. E. Fermann; 2011. "Supercontinuum generation in quasi-phaseshifted waveguides," *Optics Express* 19, 18754-18773.
- Price J.; T. Monro; H. Ebendorff-Heidepriem; F. Poletti; P. Horak; V. Finazzi; J. Leong; P. Petropoulos; J. Flanagan; G. Brambilla; X. Feng; and D. Richardson. 2007. "Mid-IR supercontinuum generation from nonsilica microstructured optical fibers," *IEEE Journal Selected Topics in Quantum Electronics* 13, 738-749.
- Sung J.; S. Lee; T. Yu; T. Jeong; and J. Lee. 2010. "0.1 Hz 1.0 PW Ti:sapphire laser," *Optics Letters* 35, 3021-3023.
- Wai P.; C. Menyuk; Y. Lee; and H. Chen. 1987. "Soliton at the zero-group-dispersion wavelength of a single-model fiber," *Optics Letters* 12, 628-630.
- Zhou B. B.; A. Chong; F. W. Wise; and M. Bache. 2012. "Ultrafast and Octave-Spanning Optical Nonlinearities from Strongly Phase-Mismatched Quadratic Interactions," *Physical Review Letters* 109, 043902.

Matteo Conforti was born in Brescia, Italy, in 1978. He received the Laurea degree and the Ph.D. degree in electronic engineering from University of Brescia, Brescia, in 2003 and 2007, respectively. He has been Research Fellow at the University of Brescia from 2004 to 2012. Currently he is Research Officer at PhLAM laboratory (CNRS-University of Lille). His main research interests include nonlinear optics and numerical methods for electromagnetism. He is the author and coauthor of over 100 refereed papers and conference presentations.

Fabio Baronio was born in Brescia (Italy). In March 2001, he received the Laurea degree (BS, MS) in Electronic Engineering (summa cum laude) from the University of Brescia, Italy. In March 2005 he received the PhD degree in Electronic and Telecommunication Engineering from the University of Padua, Italy. From March 2005 he is an Assistant Professor in Electromagnetic Fields at the University of Brescia. The scientific research activity of Fabio Baronio deals with Electromagnetism, Photonics and Telecommunications. He has authored or co-authored of over 100 refereed papers and conference presentations. He is a member of the Optical Society of America and of the IEEE Laser and Electro-Optics Society.

ON THE PROCESSING OF THE RECORDED DATA FOR THE SF6 CIRCUIT BREAKERS FROM THE TRANSFORMATION SUBSTATION 110/20/6kV CRAIOVA SOUTH

Maria Brojboiu
Virginia Ivanov
University of Craiova
Faculty of Electrical Engineering
107 Decebal Blv., 200440, Craiova, Romania
mbrojboiu@elth.ucv.ro, vivanov@elth.ucv.ro

Andrei Savescu
S.C. RELOC S.A.
109 Decebal Blv., 200746, Craiova, Romania
andreisavescu@hotmail.com

KEYWORDS

Circuit breakers, maintenance, SF6, ablation.

ABSTRACT

The assurance of the reliability for the transformation substations is one of primary goals of the manufacturers and distributors of electricity. Therefore, a proper monitoring and maintenance program is required. The circuit breaker is a complex device, subject to the thermal and mechanical stresses during the normal or fault currents switching. The substations are frequently equipped with SF6 circuit breaker. The circuit breaker components subjected to the thermal stresses are the main contacts which suffer electrical erosion and the nozzle which is subjected to the ablation process. The ablation process appears because of the energy radiation which is transferred from the electric arc. As a result of the ablation, the nozzle geometry, the gas pressure and the electric withstand are changed. Based on the recorded data from the transformation substation 110/20/6kV Craiova South, the mass loss from nozzle, the admissible number of disconnections and the throat nozzle diameter are computed. The graphical representations highlight the impact of the interrupted current, of the arcing time and of the integral I_2t over the mentioned ones. Consequently, if the switching process and the time arc value are controlled, the thermal wear can be limited and the equipments users may provide a maintenance program with minimal costs and an increasing of the lifetime.

INTRODUCTION

The circuit breaker is one of the most important and complex equipment from the medium and high voltage electric substations having the switching functions of the electrical circuit in the normal or fault conditions. Depending on the thermal and dynamic stability of the electrical equipment from stations, the commutation must be carried out in a prescribed period of time. Consequently, the failure or decommissioning of circuit breaker has undesirable effects on the operation of the power station, thereby providing a program of monitoring, diagnosis and maintenance of the circuit

breaker it is absolutely necessary. The program of monitoring, diagnosis and maintenance has the purpose to increase the lifetime of the equipment and to reduce the operation and maintenance costs. From the maintenance costs of the electrical stations, a 40% are dedicated to the circuit breaker maintenance, meanwhile a 60% are dedicated to the general revisions (Milthon S. et all, 2005). Therefore, the predictive maintenance systems based on the continuous monitoring of the circuit breaker lead to the significantly reducing of the costs. The predictive maintenance system has the advantage to be carried out during the operation of the equipment. A large number of references in the field are dedicated to the analysis of the functioning and monitoring of the circuit breakers (Milthon S. et all, 2003), (Richard, T., 2004), (Thanapong, S. 2006).

The power stations from Craiova South are equipped with oil circuit breaker or, becoming frequently after upgrading, with SF6 circuit breakers. The use of this gas has reduced the frame sizes and increase performance s of switching. The medium voltage SF6 circuit breakers are designed of the self blast principle. This type of circuit breaker generates a gas flow by means of a piston and cylinder attached to the moving contact. When the circuit breaker is in close position the gas pressure from the puffer cylinder is equal to the pressure of filling gas. During the disconnecting operation, the SF6 gas is compressed in the cavity between puffer cylinder and the piston. The switching arc occurs between the stationary contact and the moving contact and it is develops inside of the blowing convergent- divergent nozzle from PTFE with a lower thermal conductivity. A successful current disconnecting depends on the interaction between the switching arc, the radiated energy from arc, the ablation of the nozzle material and the pressure of the gas flow. During the period of the current disconnection, in normal or fault regime, occurs the thermal wear of the circuit breaker components which are in contact with the switching arc (Richard, T., 2004), (Bang, H, 2012), (Bogatyрева, N, 2013),(Muratovic, M, 2013), (Weizong W.I., 2013). The components which are directly exposed to the radiative or conductive energy transferred from the switching arc are the electrical contacts and the blowing nozzle. As a result of the thermal wear, after one operation time or

after a cumulative number of disconnected currents, these components must be replaced. In the reference (Brojboiu, M. et al, 2013) the aspects of the contact electro erosion are presented and mass loss from the contacts because of the thermal erosion is computed. It is well known that there is an admissible limit of the thermal wear beyond which the circuit breaker operation cannot be assured and this fact requires the replacement of the used components. Having in view the thermal wear of the circuit breaker, the on line monitoring of the disconnections number, the arcing time and the disconnected current values allow the estimation of the circuit breaker condition and therefore a maintenance plan can be set out.

Concerning the nozzle wear, during the operation, because of the electric arc presence, the ablation phenomenon of the nozzle material occurs. The electric energy arc is mainly absorbed by SF6 gas. An important part of this energy is absorbed by the contacts and nozzle. The energy absorption produces heating, melting and material vaporization, this being the main cause of the thermal wear. Following the ablation phenomenon occurs the increasing of the nozzle throat diameter and consequently the changing of the gas flow. In the same time, the mixing of the PTFE vapors (C2F4) with the SF6 gas appears. The influence of the vapors over the dielectric breakdown of the hot was analyzed in the reference (Weizong W.I., 2013) The PTFE vapors modify the properties of the quenching arc medium. Therefore, the quenching of the electric arc depends on the ablation intensity of the nozzle material. Consequently, the nozzle ablation has a significant influence over the composition of the gas or residual plasma between the main contacts, after the arc quenching. The withstand voltage of the gas - C2F4 vapors mixture is reduced in comparison with the one of the cold gas. The values of the critical electric field for various percentages of the PTFE vapors mixed with gas at a gas pressure of 0.40MPa were experimentally determined. Due to this fact, the re ignition of the electric arc can occurs and therefore a disconnection failure can happens. At the same time, the controlled ablation of the nozzle can produce the gas overpressure associated with the movement of a mechanical piston (Cae-Yoon B., et al, 2006). The severity of the thermal wear or the nozzle ablation depend on the amplitude of the disconnected current value, the electric arcing time, the integral I^2t , the constructive solution of the circuit breaker and the recovery voltage amplitude arc (Richard, T., 2004). There are a large number of parameters that should be monitored to evaluate the circuit breaker condition during the operation time. Concerning the main contacts of the circuit breaker, the monitoring assumes the evaluation of the thermal erosion, electric arc duration, cumulative disconnected currents and the contact resistance. The monitoring of the nozzle implies the supervision of the inner diameter of the nozzle throat, the increasing pressure in the puffer chamber and the critical electric field value.

THE PROCESSING OF THE RECORDED DATA

The medium voltage SF6 circuit breaker, whose data were recorded (Savescu A. 2013), is Fluarc FG3 type, is installed in the transformer substation 110/20/6kV Craiova South - Ghercesti. The main rated values of this circuit breaker are: rated voltage 24kV, rated current 1250A, rated breaking capacity 25kA.

The data were recorded using the protection system F650 General Electric and can be used for all equipment from transformation substations (oil or SF6 circuit breakers). Such protection system is able to store the last 20 events of the protected equipment. For circuit breakers, a recording contains the collected data in a range of time between one to two seconds. On this range of time, the currents, the voltages and the operating state of the circuit breaker (connected/disconnected/RAR) are graphically represented on every moment of time. The recorded data can be visualized by means the SIGRA4 program which is a part of the DIGSI software.

The recording of variation in time of the interrupted current is shown in Figure 1. The current values are acquired from the secondary of a current transformer with transformation ratio 200/5. The arcing time is measured from the moment of time when the current falls below the value of 1A, at this value of arcing time, the post-arc currents occur between the main contacts of the circuit breaker. In the recording from figure 1, the time duration of the arc is of 24,3ms.

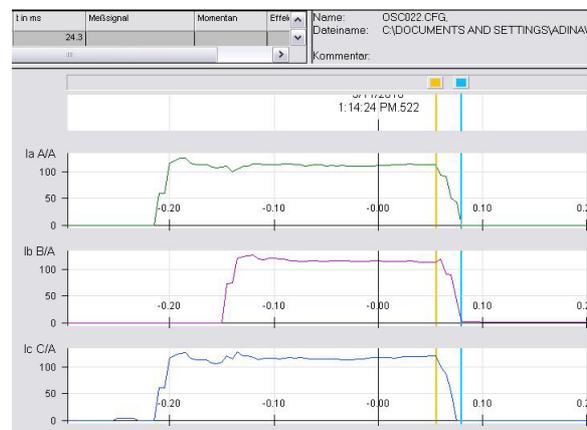


Figure 1: The Recorded Interrupted Currents vs. Time – L1,L2, L3 Phases of the SF6 Circuit Breaker

In the Figure 2 the recording of the variation in time of the voltages on the three L1,L2, L3 phases is shown. The recorded voltages are corresponding to the recorded interrupted currents. In this case, the transformation ratio of the voltage instrument transformer is of 1000/5. That means the voltage value in the moment of occurrence of the current interruption is about 22kV. The recorded interrupted currents values on three phases of the circuit breaker and the arcing time values are shown in Table 1.

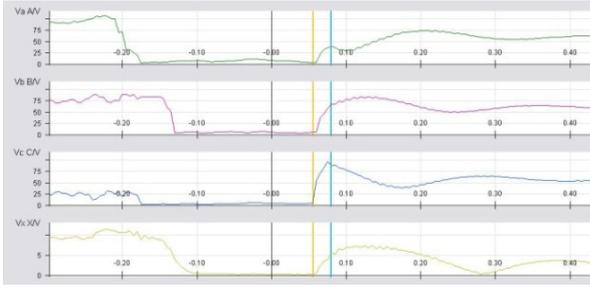


Figure 2: The Recorded Voltages vs. Time – L1,L2, L3 Phases of the SF6 Circuit Breaker

Table 1: Recorded Values: Interrupted Currents, Arcing Time / Computed Values: Mass Loss, Integral I2t

No.	I_{L1} [A]	I_{L2} [A]	I_{L3} [A]	t_a [ms]	m_{L1} [mg]	$I^2t \cdot 10^5$ [A2s]
1	4200	4560	4360	27.3	500.0933	5.0943
2	4320	4400	4480	27.3	514.3816	5.3895
3	5000	4480	4320	27.4	597.5299	7.2471
4	1316	6.8	1332	22	178.5367	0.3548
5	2228	2216	14	32.8	318.7338	1.5506
6	896	904	6.8	25.7	146.8912	0.2118
7	832	852	25.6	49.2	100.4339	0.3459
8	592	7.2	4.8	27.3	70.4893	0.1012
9	1076	5.6	9.2	31.3	143.7840	0.3490
10	1088	7.2	5.6	30.3	146.8912	0.3551
11	728	7.2	5.2	26.3	83.5076	0.1455
12	2192	6.8	2200	33.3	318.3639	1.5330
13	2160	19.6	2164	27.5	259.0750	1.3573
14	2016	9.68	2020	27.3	240.0448	1.1737

The recorded data using the SIGRA 4 program must be combined and processed in order to make conclusions on the operation system and the circuit breaker state. That could help to increase the reliability of the electrical station. Because the circuit breakers must interrupt the fault currents throughout lifetime, their aging is due to the thermal and mechanical stresses of the blowing nozzle and the main contacts between the electric arc occurs. The thermal wear or cumulative electroerosion of the main contacts is used as criterion to evaluate the electrical endurance of the circuit breaker. Additionally, for the SF6 circuit breakers is very important to assess the cumulative wear of the blowing nozzle. Accordingly to (Thanapong, S. 2006), the total admissible electroerosion value of the main contacts and of the nozzle is depending on two parameters: the maximum breaking current $I_{s_{max}}$ and the admissible number of disconnections of the fault currents N_{adm} . In the works (Brojboiu, M. et all, 2013), (Savescu, A.,2013) the computation of the mass loss from the main contacts depending on the recorded data in the oil circuit breaker from medium voltage substation Craiova is presented. The recorded data on the SF6 circuit breaker from the transformation substation 110/20/6kV Craiova

South has been processed in order to observe the dependencies between the mass loss from PTFE nozzle at the fault currents interruption and the electric arc energy, the integrals $\int i dt$ and $\int i^2 dt$. The computation of the electric arc energy was performed with formula:

$$W_a = U_a \cdot I \cdot t_a \quad (1)$$

where, the interrupted currents values and the arcing time values are presented in Table 1. The arc voltage drop U_a has been computed based on the described algorithm in (Hortopan, Gh., 1980). Knowing the maximum value of the recovery voltage as a function of maximum phase voltage, the amplitude factor k_a , first phase factor k_f , the oscillation frequency f , the following formula was deduced:

$$U_a = 0.707 \cdot u_{rmax} \cdot e^{-(T/8-\tau)} \quad (2)$$

where:

$$u_{rmax} = U_n \cdot k_a \cdot k_f \cdot \sqrt{2/3} \quad (3)$$

is the maximum value of the recovery voltage, rated voltage $U_n=24kV$, $T=1/f$, f is the oscillation frequency of the recovery voltage, τ is the time constant. In the work (Rong, M., 2005) it is estimated that 40% from electric arc power is used for the nozzle ablation, the ablation rate being around of $k_{ab}=15...17mg/kJ$. The mass loss m_{ab} as a function of the interrupted current value or the arc energy has been computed with formula:

$$m_{ab} = k_{ab} \cdot 0.4 \cdot W_a \quad (4)$$

The variation of mass loss through ablation from PTFE nozzle depending on the interrupted currents values from three phases is shown in Figure 3.

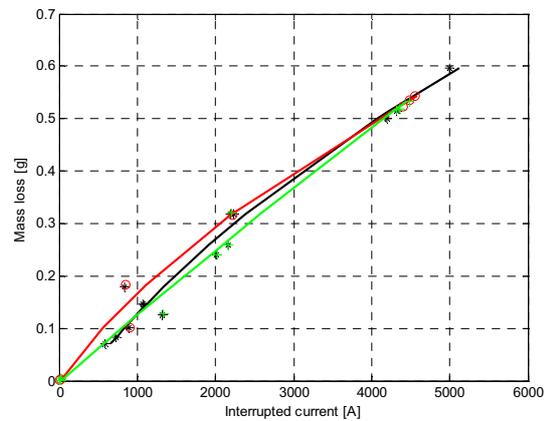


Figure 3: The Mass Loss vs. Interrupted Current- L1,L2, L3 Phases

The processing of the experimental results was performed using a Matlab application. By applying a least squares approximation method to the recorded data, using a Matlab application, it was possible to plot a continuous curve (solid line) that approximates the variation of the mass loss from nozzle depending on the interrupted current. This Matlab application has been applied to all the recorded and computed data which were processed in this work. Concerning the computation of the integral I_2t , one integration Matlab procedure has been applied. A sinusoidal variation of interrupted current was taken into account.

$$I_2t = \int_0^{t_a} (\sqrt{2} \cdot I \cdot \sin(\omega t))^2 dt \quad (5)$$

In the Figure 4 the variation of the loss mass depending on the integral I_2t is presented.

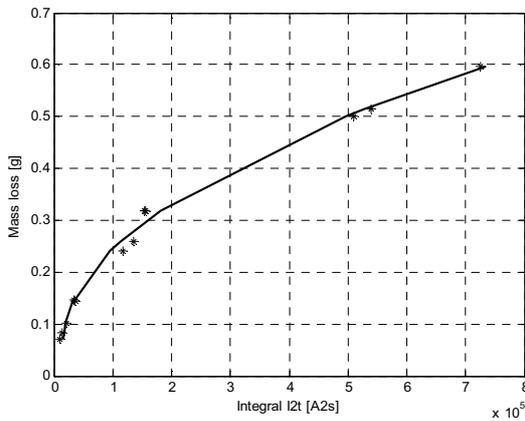


Figure 4: The Mass Loss vs. integral I_2t

From this figure, the increasing of the mass loss according to the increasing of the integral I_2t values can be noticed, as it is expected. Using the values of nozzle mass loss and for a known geometry of the nozzle, the inner radius of the nozzle throat after ablation r_a was calculated using the following formula:

$$r_a = \sqrt{r_i^2 + (m_{L1} \cdot \rho / \pi / h)} \quad (6)$$

where r_i and h are the values of the internal radius and the height of the throat nozzle respectively, measured for the circuit breaker under measurements. For the PTFE as the material of the nozzle, the density value is taken as $\rho=2200\text{kg/m}^3$.

In the Figure 5 is graphically represented the variation of the ratio of the inner diameter after ablation to the initial inner diameter, depending on the amount of the mass lost from the throat nozzle.

The large values of the ratio for reduced values of the mass loss can be observed. The experimental determinations carried out in reference (Thanapong, S.

2006) allow the establishing of the one empirical formula in order to estimate the limit mass loss depending on the admissible number of disconnections.

$$M_{\text{lim}} = 85.86 + 205.94 \cdot N \quad (7)$$

The admissible number of disconnections (the allowable number of disconnecting operations) is computed as a function of the ratio between the interrupted current and the maximum breaking current of the circuit $k=I/I_{\text{scmax}}$, $I_{\text{scmax}}=25\text{kA}$, using the following formula (Thanapong, S. 2006):

$$N_{\text{adm}} = 4.4 \cdot k^{-1.03245} \quad (8)$$

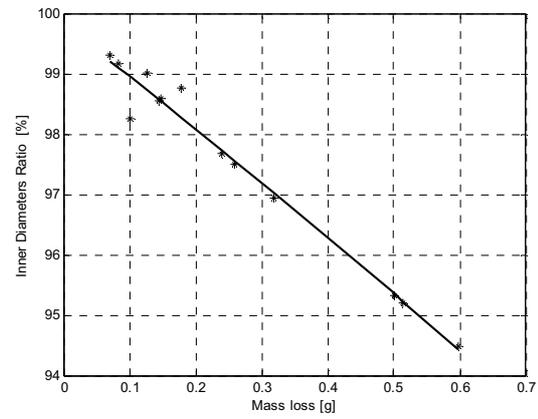


Figure 5: Inner Radius Ratio [%] vs. Mass Loss [g]

In the Figure 6, the variation of the admissible number of disconnections depending on the ratio k , for the recorded current values of the L1 phase, is shown.

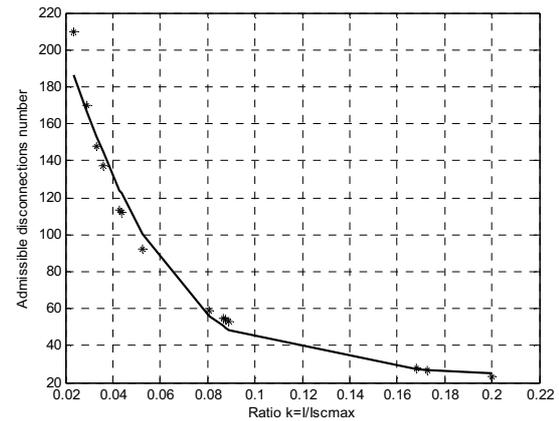


Figure 6: The Admissible Number of Disconnections vs. $k=I/I_{\text{scmax}}$

From the graphical representation, the decreasing of the admissible number of disconnections as the ratio k increases can be noticed.

CONCLUSIONS

The lifetime and the SF6 circuit breaker performances depend on the ablation intensity and the changes in the nozzle throat geometry, which are produced by the radiative and conductive energy of the electric arc. If the arcing time duration can be controlled and restricted in order to avoid the large values, the thermal wear of the contacts or the nozzle ablation can be limited. Consequently, the interval between maintenance activities can be extended.

REFERENCES

- Bang, H.; Lee, Y. S.; Ahn, H. S.; Choi, J. U.; Park, S. W. 2012. "A Study on Heat Transfer by Electric Arcs and Performance Prediction in Gas Circuit Breaker". *Recent Advances in Communications, Circuits and Technological Innovation*. Paris.
- Nadezhda, B.; Bartlova, M.; Aubrecht, V.; Holcman, V. 2013. "Mean Absorption Coefficients for SF6 + PTFE Arc Plasmas", *Electrorevue*, ISSN 1213-1539, Vol.4, No.1, APRIL 2013.
- Brojboiu, M.; Ivanov V.; Savescu, A. 2013. "Concerning the Monitoring of the Electric Contacts Electroerosion of the Circuit Breakers from Medium Voltage Stations of CEZ Craiova", *Annals of the University of Craiova, Serie Electrical Engineering*, nr.37, ISSN 1842-4805, pp.124-127.
- Himanshu, J.; Anjani, P.; Ghanashyam, P. 2013. "Optimization of High Voltage Arc Assist Interrupters", *International Journal of Scientific & Engineering Research*, Vol. 4, Issue3, March-2013, ISSN 2229-5518.
- Hortopan, Gh. 1980 "Aparate Electrice", Editura Didactica si Pedagogica, Bucuresti.
- Silva, M.; Jardini, J.; Magrini, L. 2005. "On-line condition monitoring system for in-service circuit breaker", PEA. Department of Electrical Energy and Automation Engineering, Polytechnic School of the University of Sao Paulo, Brazil.
- Silva, M.; Jardini, J.; Magrini, L. 2003. "An experience in circuit breaker on-line condition monitoring system design", PEA - Department of Electrical Energy and Automation Engineering Polytechnic School of the University of Sao Paulo, Brazil.
- Muratovic, M.; Kapetanovic, M.; Ahmethodzic, A.; Delic, S. 2013. "Calculation of nozzle ablation intensity and its influence on state of SF6 gas in thermal chamber", *Solid Dielectrics (ICSD), 2013 IEEE International Conference on*, June 30 - July 4 2013, Bologna.
- Richard, T. "Controlled Switching of High Voltage SF6 Circuit Breakers for Fault Interruption", Thesis for the degree of licentiate of engineering.
- Mingzhe, R.; Qian, Y.; Chunduo, F. 2005. "Simulation of the Process of Arc Energy-Effect in High Voltage Auto-Expansion SF6 Circuit Breaker", *Plasma Science and Technology*, Vol. 7, Issue 6, pp. 3166-3169.
- Savescu, A. 2013. "The maintenance of the 20kV circuit breaker- classical and modern constructive solutions",

Thesis for the degree of licentiate of engineering, University of Craiova, Bilateral agreement for practical training with CEZ Craiova.

Suwanasri, T. 2006. "Investigation on No-load Mechanical Endurance and Electrical Degradation of a Circuit Breaker Model under Short Circuit Current Interruption", Thesis.

Weizong, W. I. 2013. "Investigation of the dynamic characteristics and decaying behaviour of SF6 arcs in switching applications", Thesis submitted to the University of Liverpool.

AUTHOR BIOGRAPHIES



MARIA BROJBOIU is currently working as Professor at the University of Craiova, Electrical Engineering Faculty, Department of Electrical Energetic and Aerospace Engineering. Before that, she worked as design engineer at the Electroputere holding the Research and Development Center. She is Doctor in Science Technique – Electrical Engineering. She teaches the courses Electrical Equipment, Electrotechnologies and Industrial Systems Engineering. She published 5 books and 92 scientific papers for different national and international conferences and symposiums.



VIRGINIA IVANOV was born in Vela, Dolj, Romania, 1963. She was graduated in Electrical Engineering at University of Craiova, Romania, in 1986 and Doctor in Electrical Engineering in 2004. From 1986 to 1998 she worked as researcher with the Researching Institute for Motors, Transformers and Electric Equipment Craiova. In 1998 she joined the Faculty for Electrical Engineering, Department of Electrical Equipment and technologies. She is concerned with research activities in monitoring and modeling of electrical equipments.



ANDREI SAVESCU. He graduated bachelor studies in 2013 at the Faculty of Electrical Engineering at the University of Craiova and currently he is attending the Master's program "Energy quality and electromagnetic compatibility in electric systems". Since July 2013 he works as design engineer at RELOC SA company from Craiova, which has as business line the maintenance and repair of locomotive as well as designing and construction activities concerning the new locomotive models.

DESIGN, SIMULATION AND TESTING OF PLANAR SPIRAL COILS FOR THE TIME-GATED INTERROGATION OF QUARTZ RESONATOR SENSORS

Mohamad Farran, Marco Baù, Daniele Modotto, Marco Ferrari and Vittorio Ferrari
Department of Information Engineering
University of Brescia
Via Branze 38, 25123 Brescia
E-mail: m.farran@unibs.it

KEYWORDS

Quartz crystal resonators, contactless electromagnetic interrogation, planar spiral coil.

ABSTRACT

A technique for contactless electromagnetic interrogation of quartz resonator sensors is proposed and validated. The technique is based on the separation in time of the excitation and detection phases, exploiting the sensing of the transient response of the resonator. Contactless operation is achieved by means of electromagnetic coupling between spiral planar coils connected to the interrogation circuit and the sensor. The typical operating frequencies of quartz resonators in the megahertz range limit the working range to short distances. Analytical modeling and numerical and finite element simulation of the geometry of the coils have allowed to extend working operating distances up to 10 cm.

INTRODUCTION

Quartz crystal resonators (QCR) are adopted as acoustic load sensors exploiting the piezoelectric properties of quartz crystals. Typical application is the employment as quartz crystal microbalance (QCM) for measurement in biochemical applications where the mass of substances to be detected can change the vibrating mass of the resonator and hence its resonance frequency (Janshoff and Steinem. 2001). More generally, a set of acoustic properties of QCRs can be affected by the load, resulting in complex changes in the resonance response (Benes et al 1995, Ferrari and Lucklum 2008).

The possibility to contactless interrogate the resonators via electromagnetic coupling is attractive for applications where cabled solutions are generally unpractical or not allowed, i. e. in enclosed environments like hermetic boxes or sealed food packages. One of the main issues in contactless measurement information is to grant the independence of the measurement on the distance between the interrogation unit and the sensor. In this perspective, adopting a resonant principle is a robust approach, being

the measurement information in the frequency of the vibrations of the resonator and hence virtually independent on the operating distance (Ogi et al. 2006, Baù et al. 2011). Either frequency-domain or time-domain approaches have been proposed for the contactless interrogation of QCRs. Frequency-domain techniques simultaneously excite and detect the resonators, measuring the impedance or a specific transfer function, but they can be affected by signal-to-noise ratio issues and in general the measurement is dependent on the distance between the measuring unit and the sensor. Time-domain techniques, as the one proposed in the present paper, typically separate in time the excitation and detection phases exploiting the transient response of the resonator. In particular, the frequency of the damped vibrations of the resonator is exploited, which in principle depends only on the mechanical parameters of the resonator. Contactless operation can be achieved adopting capacitive or electromagnetic principles, among others. Capacitive principles, though possible do not grant sufficient operating distance, due to the weakness of the electrostatic forces on operating distances of few centimeters. Contrary, an electromagnetic technique relying on the magnetic coupling between a primary coil used for both excitation and detection and a secondary coil connected to the resonator (Lucklum and Jakoby 2009, Wu et al. 2008) has been adopted. The operating frequencies of QCRs are typically in the range of tens of megahertz and, as a consequence, the electromagnetic coupling operates on short-range distances between the primary and secondary coils.

To obtain suitable interrogation distances at feasible size of the primary and secondary coils, i.e. ultimately of the sensor element, it is of critical importance the design of the coils and simulations have a prominent role in this task. Downscaling of the dimensions of the sensor could also be beneficial for sensing applications where unobtrusiveness may be of first concern. Additionally, the measuring technique proposed herein does not need on-board power supplies, because the sensor operation relies on the intrinsic piezoelectric properties of quartz crystals and on electromagnetic coupling. In this perspective, the adoption of batteries

and related issues of maintenance and/or periodic replacement can be advantageously avoided. This paper is devoted to theoretical investigations and numerical simulations of the geometry of planar coils and their coupling. The designed coils have been realized and the simulation results have been verified by experimental tests.

OPERATING PRINCIPLE

The operating principle of the proposed interrogation technique is depicted in Figure 1. It relies on the separation in time of the excitation and detection phases. The system is composed of a primary coil L_1 which can be alternately switched between a signal generator and the input of a measuring amplifier. The primary, or TX, coil L_1 is magnetically coupled to a secondary, or RX, coil L_2 which is connected to the QCR sensor. During the excitation phase the switch is in position E and the sinusoidal signal $v_e(t)$ at frequency f_e is applied to the coil L_1 . The current in L_1 couples a magnetic field on coil L_2 and hence a current i_2 circulates in the QCR sensor connected to L_2 . The current i_2 excites the QCR into vibrations at frequency f_e . The frequency f_e is chosen to be close to the thickness-shear first resonant mode f_m of the QCR sensor, even if this is not strictly required because the proposed principle relies on the detection of the free decaying response of the resonator which is independent of the excitation frequency. This is advantageous because the resonant frequency of the QCR cannot be known in advance. The ideal condition $f_e=f_m$ enhances the signal-to-noise ratio during the detection phase. When the switch is set to position D after a period of time T_E , the QCR sensor undergoes decaying harmonic oscillation to its damped resonant frequency f_{dm} . This oscillation, thanks to the piezoelectric properties of quartz, induces a current into L_2 which in turn couples to L_1 . The induced voltage $v_o(t)$ on L_1 is read by means of an amplifier. The signal $v_o(t)$ is fed to a zero-crossing detector to derive a square wave signal at frequency $f_o=f_{dm}$ measurable with a frequency counter. If the

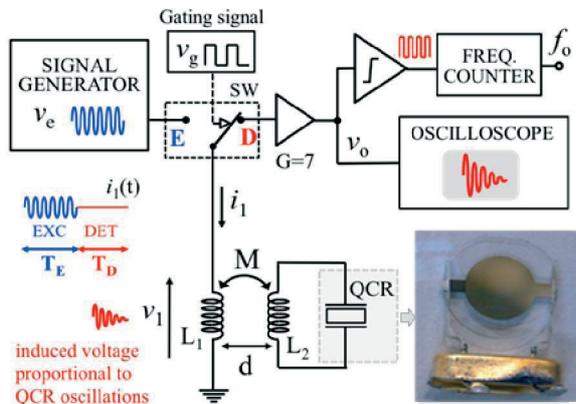


Figure 1: Schematic diagram of the operating principle.

mechanical quality factor Q of the QCR sensor is high, then $f_o=f_m$. The readout signal is sensed for the period T_D after which a new excitation phase starts. Considering high- Q QCR sensors, the voltage $v_1(t)$ sensed across the inductor L_1 during the detection phase can be approximated by (Baù et al. 2011):

$$v_1(t) = MA_m \exp(-t/\tau_m) \cos(2\pi f_{dm}t + \theta_m) \quad (1)$$

where M is the mutual inductance between the coil L_1 and L_2 , A_m and θ_m are the amplitude and phase terms which depend on the initial mechanical and electrical conditions at the beginning of the detection phase. The exponential decaying time τ_m is related to the mechanical quality factor by $Q=\pi f_m \tau_m$ while the damped resonance frequency is related to the mechanical resonance frequency by $f_{dm}=(4\pi^2 f_m^2 - 1/\tau_m^2)^{1/2}$. In (1) it can be observed that the mutual inductance M acts as a scaling factor, without affecting the frequency of the readout signal. On the other side, the value of M depends on the distance, orientation and geometry of the coils and therefore sets the maximum operating distance of the proposed interrogation principle. Hence, the design of the proper geometry of the coils is fundamental to achieve larger operative distances and it is subject to dimensional, electrical and mechanical constraints.

INDUCTIVE COUPLING

To optimize the operating distances between the primary and secondary coils, the influence of the geometrical and electrical parameters of the coils on the electromagnetic coupling have been investigated by means of theoretical methods as well as by numerical simulation.

Coil equivalent circuit

Several closed-form expressions have been proposed to approximate the inductance L of printed spiral coils (PSCs). We adopted (2) from (Mohan 1999) for square-shaped coils whose layout is shown in Figure 2 on the left

$$L = \frac{2\mu_0 n^2 l}{\pi} \left[\ln\left(\frac{2.067}{\phi}\right) + 1.078\phi + 0.125\phi^2 + 0.5 \frac{(n-1)s^2}{(\phi)^2} \right] + \frac{2\mu_0 n^2 l}{\pi} \left[0.178 \frac{(n-1)s}{nl} + 0.0833 \frac{(n-1)s(s+w)}{l^2} - \frac{1}{n} \ln\left(\frac{w+t}{w}\right) \right] \quad (2)$$

$$\phi = \frac{d_{out} - d_{in}}{d_{out} + d_{in}} \quad (3)$$

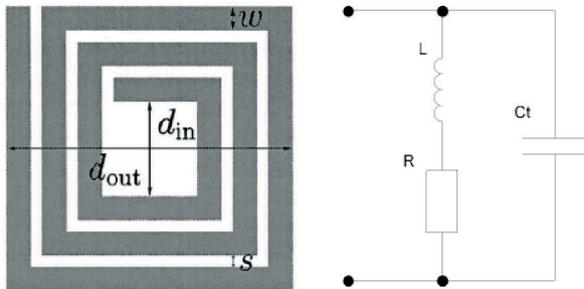
where n is the number of turns d_{out} and d_{in} are the outer and the inner diameters of the coil, μ_0 is the vacuum permeability of space. The wire cross-section is rectangular and its width and thickness are w and t , respectively. The wire sections are uniformly spaced with edge-to-edge spacing s ; φ is a parameter known as fill factor, which changes from 0, when all the turns are concentrated on the perimeter like filament coils, to 1, when the turns spiral all the way to the center of the coil.

The right panel of Figure 2 shows the equivalent resonant circuit of the PSC. The series resistance is calculated using the following expression

$$R = \frac{\rho l}{w\delta(1 - e^{-\frac{t}{\delta}})} \quad (4)$$

$$l = 4nd_{out} - 4nw - (2n + 1)^2(s + w) \quad (5)$$

where l is the length of the conductive trace, ρ is the resistivity of the conductive material and δ is the metal skin depth. It must be emphasized that the AC-resistance is also affected by proximity effects caused by the currents that are flowing through one or more other nearby conductors, such as within a closely wound coil of wire. The total capacitance is affected by the presence of two insulating materials: one is the air filling the gap between adjacent traces and the other is the PSC substrate, which could be ceramic, polyimide, or FR4. In literature there are many analytical approximations for the parasitic capacitance, but the estimation accuracy can be insufficient, therefore it becomes necessary to use the available simulation tools to validate the design result. In this study, 3-D numerical simulations were used for an accurate prediction of the parasitic elements and for the final tuning of the proposed PSCs. Even though many fabrication processes allow using multilayer conductors, we limited our design to single layer PSCs because the parasitic capacitance for multilayer PSCs is much larger than for single layer PSCs and this can significantly reduce the self-resonance frequency (SRF).



Figures2: Geometrical parameters of a square-shaped printed spiral coil. Schematic drawing (on the left) and equivalent resonant circuit (on the right).

Coils mutual inductance

The derivation of the general mutual inductance formula for planar structures starts from the mutual inductance between two circular air cored loops whose axes are parallel; by using Maxwell's equations, one can obtain

$$\begin{aligned} M(a, b, \gamma, d) &= \pi\mu_0\sqrt{ab} \int_0^\infty J_1\left(x\sqrt{\frac{a}{b}}\right) J_1\left(x\sqrt{\frac{b}{a}}\right) \\ &\times J_0\left(x\frac{\gamma}{\sqrt{ab}}\right) \exp\left(-x\frac{d}{\sqrt{ab}}\right) dx \end{aligned} \quad (6)$$

where a and b are the radii of the two coils, d is the separation between the coils and γ is distance between their axes. J_0 and J_1 are the Bessel functions of the zeroth and first order.

This expression does not contain the radius R of the coil's wire since it is assumed that the ratios R/a and R/b are sufficiently small (Zierhofer and Hochmair 1996). For perfectly aligned coaxial coils, where $\gamma = 0$, (6) can be simplified to the Neumann's expression treated in numerous text books (Jow and Ghovanloo 2007):

$$M(a, b, d) = \frac{2\mu_0\sqrt{ab}}{\alpha} \left[\left(1 - \frac{\alpha^2}{2}\right) K(\alpha) - E(\alpha) \right] \quad (7)$$

where $K(\alpha)$ and $E(\alpha)$ are complete elliptic integrals of the first and second kind, respectively and where

$$\alpha = 2\sqrt{\frac{ab}{(a+b)^2 + d^2}} \quad (8)$$

Since a PSC can be considered as a set of concentric single turn coils with shrinking diameters and connected in series, we can use (5) or (6) to calculate the overall M of two PSC in parallel planes: the mutual inductance value can be found by summing the values of the partial mutual inductances between every turn of one coil and all the turns of the other coil

$$M_{tot} = g \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} M_{ij}(a_i, b_j, \gamma, d) \quad (9)$$

where g is a factor dependent on the shape of the PSC and in (Jow and Ghovanloo 2007) they found empirically that $g \approx 1.1$ for a square-shaped PSCs.

In this study, the primary coil (TX) and the secondary coil (RX) are separated by a distance d which is typically in the range of a few centimeters. The coils are loosely coupled and are characterized by extremely low coupling coefficients ($k \leq 0.01$) (Fotopoulou and Brian 2011). As a matter of fact, this happens when the desired separation (d) between the TX and RX coils is larger than the coil dimensions as is the case in most of radio frequency identification (RFID) applications.

For this separation values, we can take advantage of the properties of the magnetic field at large distance and we can assume a homogeneous magnetic field. Under this

hypothesis, the mutual inductance M between two distant square spiral coils in coaxial configuration can be readily calculated:

$$M = \frac{B(0,0,d)A_{RX}}{I} \quad (10)$$

where A_{RX} is the area of the RX coil and I is the current that excites the TX coil. The dominant B_z magnetic field component at the center of the RX coil can be obtained by integrating Biot-Savart's law around the TX coil having side length L

$$\vec{B}(0,0,z) = \frac{\mu_0 I}{2\pi \left(\frac{z^2}{L^2} + \frac{1}{4}\right) \sqrt{z^2 + \frac{L^2}{2}}} \hat{z} \quad (11)$$

When the RX coil is distant from the reader antenna (further than a side length L , such that $z \gg L$), the magnetic field falls off rapidly:

$$\vec{B}(0,0,z) \approx \frac{IL^2}{2\pi z^3} \hat{z} \quad (12)$$

The following expressions can be generalized for a PSC with n turns where $L=d_{in}$ and from (12) we can conclude that the optimal value for d_{in} depends on the coils relative distance, d .

We demonstrated by comparing (10) and (7) for a set of PSCs varying the distance of separation that $d > d_{out}$ is enough for considering a homogeneous magnetic field at the RX coil.

We have used equations (2)-(11) in combination with CST (Computer Simulation Technology) and COMSOL Multiphysics to find the optimal coil geometries. We have also fabricated a number of PSCs on FR4 substrate, designed through a procedure which will be explained in the next section.

Simulation of coil geometries

The self-inductance of the TX and RX coils required by the sensor circuit at the operation frequency of 4.8 MHz is 8 μ H and the desired reading distance is more than 5 cm. In order to design the coil and optimize the interrogation range we developed a MATLAB code devised for coils made of 1-D filaments wires by sweeping the parameters included in (2), (4) and (10). Furthermore, we tried to maximize the mutual inductance M by designing a TX PSC with d_{in} as large as possible according to (11). We fixed s (200 μ m), and t (35 μ m) at the minimum value permitted by the fabrication technology even if this choice may not be optimal. In (2) and (4) we swept n and d_{in} for a fixed inductance value L (8 μ H) and we selected the TX PSC ($n=6$) with a high d_{in} (69.2 mm) which we analyzed and optimized numerically with CST. Note that at high frequencies (>10 MHz) the effects of PSC parasitic capacitance becomes more significant for PSCs with larger diameters and the self-resonance frequency (SRF) moves downward quickly; w is 400 μ m to guarantee low

resistance loss ($\approx 3\Omega$) within a large frequency range including 4.8 MHz.

After selecting the geometrical parameters of the TX PSC we focus on the secondary PSC. Ideally, for a compact reading device, the d_{out} RX PSC should be comparable to the sensing passive device. The sensor transmits in the detection phase so we have to proceed as before to reach a large operation distance. Table I shows the geometries and specifications of four square-shaped PSCs that were designed and then fabricated using 35 μ m thick copper on FR4 substrates. PSC2~4 were used as RX PSC and PSC1 as RX-TX PSC. Figure 3 shows the trend of the mutual inductance versus the distance of separation for fabricated coils in Table I calculated numerically by the simulations performed in COMSOL Multiphysics. The optimization of the mutual coupling considered not only the aligned coaxial coils but also the presence of misalignment between the coil axes. Figure 4 shows the calculated mutual inductance values for a distance $d=6$ cm as a function of the displacements in the (x,y) plane perpendicular to the coil axis (along z).

EXPERIMENTAL VALIDATION

Planar coils for the experimental test have been fabricated by milling a FR4 substrate with 35 μ m-thick

Table I: Specifications of the PSCs used in measurements

Parameter	PSC1	PSC2	PSC3	PSC4
Shape	Square	Square	Square	Square
d_{out} (mm)	76	34.3	31.2	30
d_{in} (mm)	69.2	17.9	10.1	1.6
n(turns)	6	14	18	24
w(μ m)	400	400	400	400
s(μ m)	200	200	200	200
L(μ H)	8.027	8	8.35	8.5
R(Ω)	3	2.52	3.1	2.65
SRF(MHz)	22.4	48	50	54

copper layer. The geometric properties of the coils are the same reported in Table 1. In particular, two coils corresponding to PSC1 in Table 1 have been considered. A picture of the coil is shown in the inset of Figure 5. By means of a HP4194A impedance analyzer the mutual inductance M between the two coils at various distances has been measured. Figure 5 compares the values of the mutual inductance measured at 6 MHz with the values predicted by COMSOL simulations. Additionally, the system described in Fig. 1 has been tested adopting a QCR with resonance frequency of about 4.8 MHz as the sensor. Multiple coils have been fabricated according to the geometrical parameters of Table 1. In particular, a PSC1 coil has been connected to the interrogation circuit during all the experimental tests, while the sensor has been connected alternatively to all the PSC1~PSC4 coils.

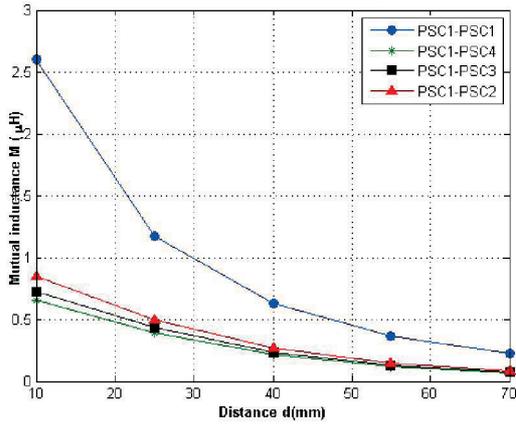


Figure 3: Mutual inductance versus the separation distance d for the fabricated coils of Table I (values calculated by means of COMSOL Multiphysics).

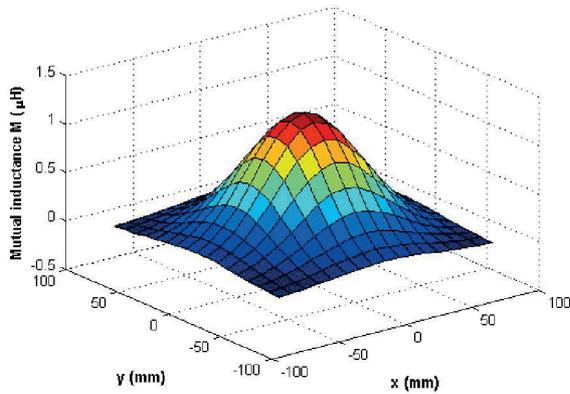


Figure 4: Mutual inductance computed for PSC1-PSC2; PSC1 is centered at the origin and is parallel to the (x,y) plane and PSC2 moves in the (x,y) plane at the separation distance $d=6$ cm.

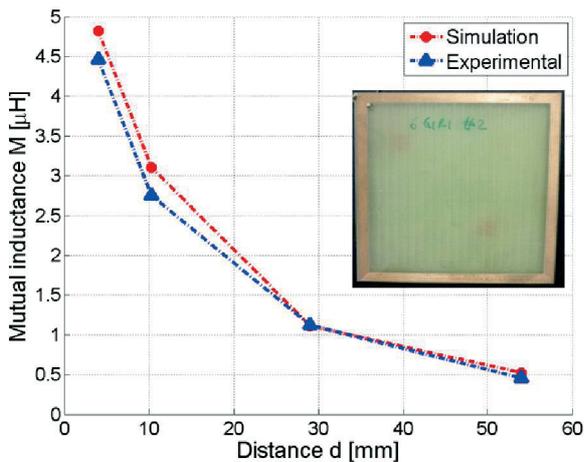


Figure 5: Comparison between the simulated and measured values of the mutual inductance M for the PSC1 coil.

The transient response of the sensor has been acquired by means of a digital oscilloscope Agilent

MSO-X3014A starting from a prescribed delay after the gating signal has switched to the detection phase. The acquired signal FFT has been calculated and the amplitude of the peak in correspondence of the sensor resonance frequency has been plotted as a function of the interrogation distance d , as shown in Figure 6. As it can be observed, the trend of the experimental points for each of the coil pair follows the trend of the simulated values of the mutual inductance M of Figure 3, confirming that M acts as a scaling factor on the response of the sensor.

CONCLUSIONS

A technique for the contactless electromagnetic interrogation of QCR sensors which alternatively excites the sensor and detects the transient response, has been proposed. The combination of the time-gated interrogation technique with the adoption of a resonant sensing principle is robust with respect to the distance between the interrogation and sensor unit because the measurement information is carried by the resonance frequency of the resonator. The typical operating frequencies of QCRs are in the megahertz range, confining the operation of the proposed technique to short-range distances. To enhance performance in terms of operating distances, we optimized the geometries of printed spiral coils adopting a semi-empirical method, using a combination of theoretical results (1-D filament) and 3-D numerical simulations.

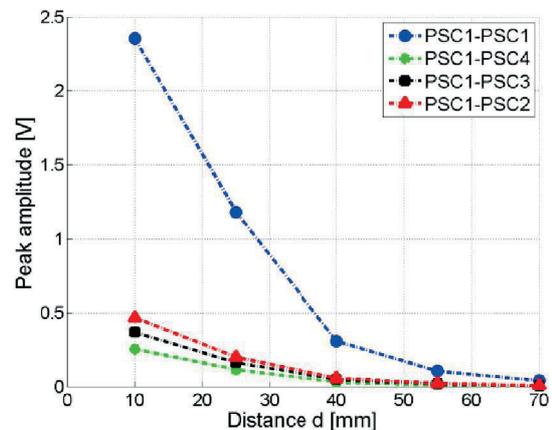


Figure 6: Amplitude of the FFT magnitude peak in correspondence of the resonance frequency, measured by means of the interrogation circuit for a subset of the coil pairs of Table I

This method takes into account the effects of the parasitic elements (resistance and capacitance) on the operative band and mutual inductance of the designed coil. The obtained experimental results confirm the predictions of the lumped element equivalent model in Figure 1, demonstrate the effectiveness of the approach and show an optimized working distance of up to 10 cm in the case PSC1-PSC1. The proposed system can be exploited for the measurement of physical/chemical quantities affecting the resonant response of QCR sensors.

REFERENCES

- Benes E., M. Gröschl, W. Burger and M. Schmid. 1995. "Sensors based on piezoelectric resonators". *Sens. Actuators A*, 48, pp. 1–21.
- Baù M., M. Ferrari and V. Ferrari, E. Tonoli. 2011. "Electromagnetic contactless interrogation technique for quartz resonator sensors". *IEEE Sensors 2011*, pp. 1297-1300.
- Ferrari V. and R. Lucklum. 2008. "Overview of acoustic-wave microsensors". In A. Arnau (Ed.), *Piezoelectric Transducers and Applications (2nd ed.)*, Springer, pp. 39-62.
- Fotopoulou K. and W.F. Brian. 2011. "Wireless power transfer in loosely coupled links: Coil misalignment model", *IEEE transactions on magnetics*, vol. 47, pp. 416-430.
- Janshoff A. and C. Steinem. 2001. "Quartz crystal microbalance for bioanalytical applications". *Sensors Update*, 9, pp. 313–354.
- Jow U. and M. Ghovanloo 2007, "Design and optimization of printed spiral coils for efficient transcutaneous inductive power transmission". *IEEE transactions on biomedical circuits and systems*, vol. 1, No.3, pp. 193-202.
- Lucklum F. and B. Jakoby. 2009. "Non-contact liquid level measurement with electromagnetic-acoustic resonator sensors". *Meas. Sci. Technology*, 20,p. 124002 (7pp).
- Mohan S. S. 1999, "The design, modelling and optimization of ON-CHIP inductor and transformer circuits". PHD thesis, Stanford University.
- Ogi H., K. Motoshisa, T. Matsumoto, K. Hatanaka, M. Hira. 2006. "Isolated electrodeless high-frequency quartz crystal microbalance for immunosensors". *Analytical Chemistry*, 78, pp. 6903–6909.
- Wu W., D.W. Greve, I.J. Oppenheim. 2008. "Inductively coupled sensing using a quartz crystal microbalance". *Proceedings of IEEE International Ultrasonics Symposium*, pp. 1018–1021.
- Zierhofer C.M. and E.S Hochmair. 1996, "Geometric approach for coupling enhancement of magnetically coupled coils". *IEEE transactions on biomedical engineering*, vol. 43, No.7, pp. 708-714.

AUTHOR BIOGRAPHIES

MOHAMAD FARRAN was born in Nabatieh, Lebanon and he obtained his master degree in telecommunication engineering from the Università di Brescia in 2012. He is currently working toward the Ph. D. degree in Electronic Instrumentation. His research activities include planar antennas, RFID systems and contactless resonant sensors. His e-mail address is: m.farran@unibs.it.

MARCO BAU' was born in Castiglione delle Stiviere, Italy. He obtained the laurea degree in Electronic Engineering 2005 and the Ph. D. degree in Electronic Instrumentation in 2009. His research activities deal with the investigation of techniques for the contactless interrogation of resonant sensors, the design of MEMS sensors and related front-end circuits and energy harvesting techniques and circuits. His e-mail address is: marco.bau@unibs.it.

MARCO FERRARI was born in Brescia, Italy, in 1974. In 2002, he obtained the Laurea degree Electronics Engineering degree at the University of Brescia. In 2006 he received the Research Doctorate degree in electronic instrumentation at the same university. Since 2007 he has become an assistant professor at the Department of Electronics for Automation of the University of Brescia. His research activity deals with the energy conversion via the piezoelectric effect for powering autonomous microsystems and sensors for physical and chemical quantities with the related signal-conditioning electronics. In particular he is involved with piezoelectric acoustic-wave sensors in thick-film technology, design of oscillator circuits and frequency-output signal conditioners. His e-mail address is: marco.ferrari@unibs.it.

DANIELE MODOTTO was awarded the Laurea degree (cum laude) in Electronic Engineering and the Ph.D. degree in Electronics and Telecommunications Engineering from the Università di Padova, in 1996 and 2000 respectively. From 2000 to 2001 he was with the Department of Electronics and Electrical Engineering, University of Glasgow, where he was involved in nonlinear wavelength conversion modeling. Since November 2001 he has been working at the Università di Brescia and since November 2002 he is an Assistant Professor of Electromagnetic Fields. His scientific contributions cover three main areas: microwave antennas, nanophotonics and microstructured optical fibers. His e-mail address is: daniele.modotto@unibs.it.

VITTORIO FERRARI was born in Milan, Italy, in 1962. In 1988, he obtained the Laurea degree cum laude in Physics at the University of Milan. In 1993 he received the Research Doctorate degree in Electronic Instrumentation at the University of Brescia. He has been an assistant professor and an associate professor at the Faculty of Engineering of the University of Brescia until 2001 and 2006, respectively. Since 2006 he has been a full professor of Electronics. His research activity is in the field of sensors and the related signal-conditioning electronics. Particular topics of interest are acoustic-wave piezoelectric sensors, microresonant sensors and MEMS, energy harvesting for autonomous sensors, oscillators for resonant sensors and frequency-output interface circuits. He is involved in national and international research programmes, and in projects in cooperation with industries. He serves in international panels, conference committees and boards in the field of sensors and electronic instrumentation. His e-mail address is: vittorio.ferrari@unibs.it.

APPLICATIONS OF THE GRAPH THEORY FOR OPTIMIZATION IN MANUFACTURING ENVIRONMENT OF THE ELECTRICAL EQUIPMENTS

Virginia Ivanov
Maria Brojboiu
Sergiu Ivanov
University of Craiova
Faculty of Electrical Engineering
107 Decebal Blv., 200440, Craiova, Romania
vivanov@elth.ucv.ro, mbrojboiu@elth.ucv.ro, sergiu.ivanov@ie.ucv.ro

KEYWORDS

Graph theory, Hamiltonian path, electrical equipments.

ABSTRACT

Depending on user requirements, manufacturing systems dedicated to electrical equipment must produce a wide range of products. The transition from the manufacturing an assortment of product to another involves additional costs which are necessary to adjust the manufacturing system state to the new technology. The manufacturing optimization requires the launching in fabrication of the assortments of products in a predetermined sequence in order to minimize the cost of changing the technical condition of the system and its adaptation to the technological specificity of the new sort. The graph theory can be successfully used in order to optimize the launching of different type of products and the optimal paths which allow minimal costs. Therefore, one can solve the problem of determining the optimal Hamiltonian path from the point of view of minimal time for scanning a certain path. Several applications of optimum Hamiltonian path will be then presented in this paper. They use either the Chen algorithm or depth-first one, being integrated in the same software application.

INTRODUCTION

The peculiarities of design and manufacture of the electrical equipment are a consequence of many functions that must be fulfilled (switching, protection, instrumentation, amplification etc). A very large range of rated values such as: the rated voltage range from 1 kV to hundreds of kV and the rated current range from hundreds of amperes to tens of kA, have a major influence also.

The wide range of electrical equipment that are manufactured in a dedicated production system includes circuit breakers, contactors, fuses, disconnectors, surge arresters, switches, metal clad switchgear or instrument transformers etc. The constructive solution design of each type of equipment depends on the function that it

meets, the rated voltage level and the specific environment requirements.

In this context, the design and manufacturing of the electrical equipment require the use of a wide range of conductors, insulating, magnetic materials, and so on. Accordingly, the manufacturing is characterized by the specific technologies for each type of product, specific machinery or equipment, dedicated tools, measurement and control device which are diversified and tailored to the specific measurements. Providing by the user of the reliable requests of the products and the high efficiency of the manufacturing technologies in order to reduce the costs is one of the major objectives of equipment manufacturers.

The efficiency of the manufacturing technologies is accomplished through the use of group technologies and implementing of systematization criteria of the constructive solutions. However, under these circumstances, the transition from the manufacturing an assortment of a product to another involves additional costs which are necessary to adjust the manufacturing system state to the new technology. This means, for example, new settings of machinery, new tools, new materials acquisition, changing instrumentation and control device, new operators during the production, using of other internal transportation system etc. Therefore, the manufacturing optimization requires the launching in fabrication of the assortments of products in a predetermined sequence in order to minimize the cost of changing the technical condition of the system and its adaptation to the technological specifics of a new sort.

Consequently, an optimal sequence of electrical equipment assortments to be launched in manufacturing process allows the optimization by minimizing of the costs.

Among the amelioration methods which can be applied in the manufacturing environment, the graph theory can be successfully used in order to optimize the launching of different type of products and the optimal paths which allow minimal costs.

Many times, the number of possibilities is very large and then, the optimization problem gets another dimension: the time of selection. For these cases, a formalized form

is useful in order to be implemented in a numerical algorithm run by a computer. Even more, it can define another concern: the methods' optimization that means to find that methods which give, on the shortest way, the optimal solution (Abrudan 1980).

The interest for graph theory has grown in the last decades, when it was applied for different problems in areas as economy, sociology, psychology, engineering etc.

The complex systems and situations can be represented by graphs. Its clearly highlight all the aspects of particular states.

A graph G is completely defined by the nodes array X and by the edges array U .

Mathematically, a graph (G) is:

$$G = (X, U), \quad (1)$$

where,

$$X = \{x_1, x_2, \dots, x_n\}, \quad (2)$$

is the nodes array and:

$$U = \{(x, y) | x, y \in X\}, \quad (3)$$

is the edges array.

If we look a graph as the image of a system, the nodes are the system's components and the edges (x_i, x_j) are interdependencies between components. Even a component x_i does not directly influences the component x_j , it can influences by the way of other components. In this case there is a chain of intermediate components $\{x_1, x_2, \dots, x_k\}$. Each component directly influences the next component and finally influences x_j . Each edge (x_i, x_j) signifies that the system can directly switch from node x_i in node x_j .

The minimum cost – optimizing criteria of the manufacturing environment

The optimization refers to the process economy. The objective can be assumed to be the minimization of the adaptation effort for variable production task. An important component of the efforts is represented by the transition costs generated by the switching between different technologies, depending on the manufactured products.

Optimizing the system means to optimize the operation of the system and mainly the transition costs among the life of the manufacturing system.

The transition costs can be mathematically described by the transition cost matrix. The manufacturing technology corresponding to the product P_i is represented by the state S_i of the system. When the manufacturing of the product P_i begins, the system being in the state S_j , a transition cost is generated for the transition from S_j to S_i . This transition cost is denoted as c_{ij} .

Generally, the transition cost matrix (C) can be expressed as:

$$\begin{matrix} & \begin{matrix} [S_1 & \dots & S_j & \dots & S_m] \\ \Downarrow & \dots & \Downarrow & \dots & \Downarrow \end{matrix} \\ \begin{matrix} [P_1] \\ \vdots \\ P_i \\ \vdots \\ [P_m] \end{matrix} \Rightarrow & C = & \begin{bmatrix} 0 & \dots & c_{1j} & \dots & c_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{i1} & \dots & c_{ij} & \dots & c_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{m1} & \dots & c_{mj} & \dots & 0 \end{bmatrix}, \end{matrix} \quad (4)$$

where:

$$S = \{S_1, S_2, \dots, S_j, \dots, S_m\}, \quad (5)$$

is the array of the states of the manufacturing system.

$$P = \{P_1, P_2, \dots, P_j, \dots, P_m\}, \quad (6)$$

is the array of the products types which will be manufactured by the analyzed system.

The elements of the transition costs matrix are as follows:

$$c_{ij} = \begin{cases} c_{ij} > 0, & \text{for } i \neq j \\ c_{ij} = 0, & \text{for } i = j \\ c_{ij} \neq c_{ji} \end{cases}. \quad (7)$$

Hamiltonian paths

For a manufacturing system used for a set of products, it must be revealed the optimal sequence of production launches which minimizes the transition costs. The problem can be solved by applying the graph theory. The solution will be the optimal Hamiltonian path.

One of the most popular economic problems is the optimal Hamiltonian path from the point of view of minimal time for scanning a certain path. The minimum time is equivalent to the shortest path which touches once each node. In addition, the final state must be the same as the initial one. The literature signals more algorithms, precise or heuristic, which can give a satisfying solution of the optimal Hamiltonian path without significant delay.

Chen Algorithm

When the Hamiltonian path must be found in a graph without circuits, the Chen algorithm can be applied. It consists in the following steps (Fiedler 1973):

1. The adjacency matrix (A) is defined. This matrix will be a $n \times n$ matrix for a system with n states:

$$A = [a_{ij}], 1 \leq i, j \leq n, \quad (8)$$

where,

$$a_{ij} = \begin{cases} 1, & \text{if the edge } (x_i, x_j) \text{ exists} \\ 0, & \text{in rest} \end{cases}. \quad (9)$$

2. The elements of the paths matrix (D) are determined:

$$D = [d_{ij}], 1 \leq i, j \leq n, \quad (10)$$

where,

$$d_{ij} = \begin{cases} 1, & \text{if at least one edge } (x_i, x_j) \text{ exists} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

3. The applicability of the Chen algorithm is checked (graph without circuits):

If there is an index i with $d_{ii} = 1$, then the graph has circuits and the Chen algorithm can't be applied.

4. If not, the vertex connectivity for each node $p(x_{i..n})$ is computed.

The vertex connectivity of a node x_i , denoted $p(x_i)$, is the maximum number of nodes which can be reached starting from the node x_i .

If, from any node x_i all the superior nodes and only these can be reached, the vertex connectivity of the node x_i is

$$p(x_i) = n - i. \quad (12)$$

5. The equality is checked:

$$\sum_{i=1}^n p(x_i) = \frac{n \cdot (n-1)}{2}. \quad (13)$$

If (13) is true, the next step is 6. Otherwise the graph does not have Hamiltonian paths and the algorithm stops.

6. The nodes are descendent ordered according to the vertex connectivity and the optimal Hamiltonian path is obtained.

„Depth-First” Algorithm

Another algorithm used for minimum Hamiltonian path determination is the “Depth-First” Algorithm.

In this strategy, the nodes are explored depending on the depth and the upper level is resumed only if the search arrives in a dead-end. The depth search is not completed for infinite trees, when cycles occur. If a constraint is defined for avoiding repeated states, the strategy can be completed. On the other hand, the strategy can achieve the solution in the most depth, without taking into account the cost and consequently, it is not optimal. If additional minimal costs constraints are added, the algorithm can be optimized.

The algorithm consists in following steps (Chin et al. 1982):

1. The adjacency matrix (A) is defined as in (8) and (9).
2. The transition costs matrix C is defined (the costs for transition from state i to state j):

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}. \quad (14)$$

The searching algorithm expands each node and explores the edges down to the final node. So, a path is stated.

In order to retrieve the rest of the paths, the algorithm returns to the previous node and tries to find another path, different by the one already stated. The algorithm ends when all the paths were checked and consequently all the Hamiltonian paths were found.

The most important advantage of the depth-first algorithm resides in the requested memory resources which a minimal due to the linear complexity. The algorithm must keep only a single path from the initial node to the current final node, together with all the unexpanded nodes which are “brothers” (have a common predecessor) with the nodes in the current path. The main disadvantage of the algorithm is that is not optimal and it is not completed if additional constraints are not considered.

APPLICATIONS FOR PREDETERMINED SEQUENCE OF THE PRODUCTS MANUFACTURING

For the complex manufacturing environments, the optimum element should be selected from a large array and consequently, the time for selection increases too much. This why, it is useful to have an algorithm capable to be run on a machine.

Following, several applications of optimum Hamiltonian path will be presented. They use either the Chen algorithm or depth-first one, being integrated in the same software application.

When the software application is launched, the user can choose the algorithm to be used: Chen or depth-first (Fig. 1).

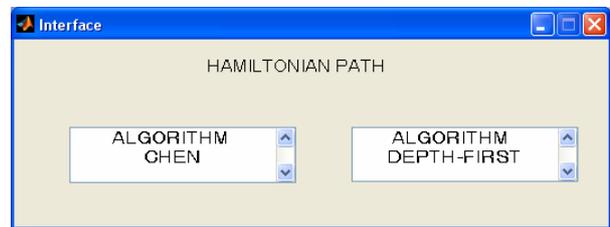


Figure 1: The Selection Algorithms

FINDING THE OPTIMUM HAMILTONIAN PATH BY USING CHEN ALGORITHM

This algorithm can be used only for oriented graphs which does not contain circuits. The returned result will be the single Hamiltonian path of a given graph.

The input data are defined (Fig. 2):

- the number of nodes which represents the number of products and the number of states respectively. For example, $n = 10$;
- the adjacency matrix, A , an $n \times n$ matrix, defined in accordance with (8) and (9):

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (15)$$

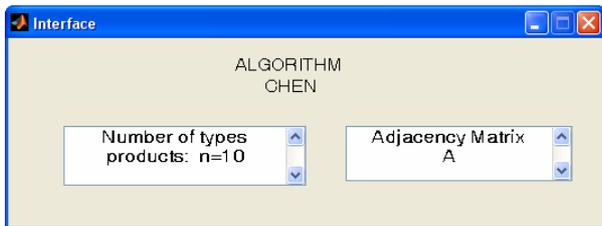


Figure 2: The Input Date of Chen Algorithm

The application determines the paths matrix, in accordance with (10) and (11):

$$D = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (16)$$

The existence of circuits within the given graph is checked. This means to check if for any i we have $d(i,i) = 1$. In this case the application ends (Fig. 3).

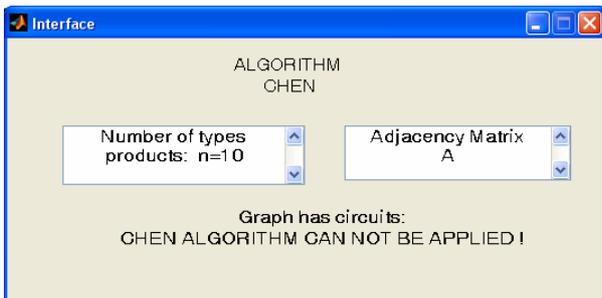


Figure 3: The Message “STOP APPLICATION”

If the graph does not contain circuits, the vertex connectivity is computed for each node based on (12) and then the condition (13) is checked. If this condition is not fulfilled, it means that there is not any Hamiltonian path (Fig.4).

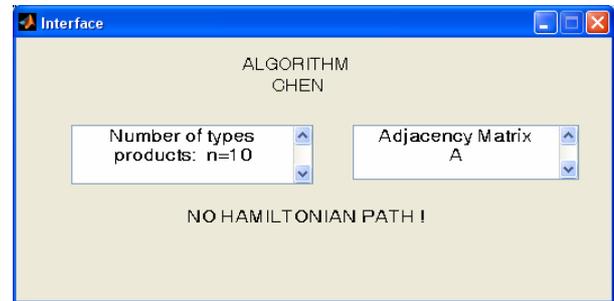


Figure 4: The Message “NO HAMILTONIAN PATH”
If the condition (13) is fulfilled, the optimum Hamiltonian path results as the sequence of scanning the graph’s nodes (Fig.5).

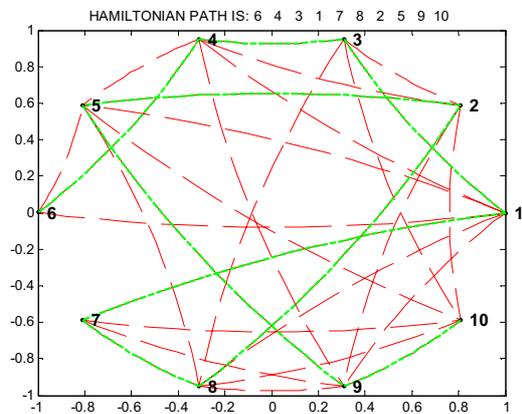


Figure 5: Optimum Hamiltonian Path

FINDING THE OPTIMUM HAMILTONIAN PATH BY USING DEPTH-FIRST ALGORITHM

The developed software application can analyze oriented, non-oriented, with or without circuits graphs. It returns the optimum Hamiltonian path from the point of view of costs (Chin et al. 1982). The input data are: the transitions costs matrix (C) and the adjacency matrix (A) (Fig.6).

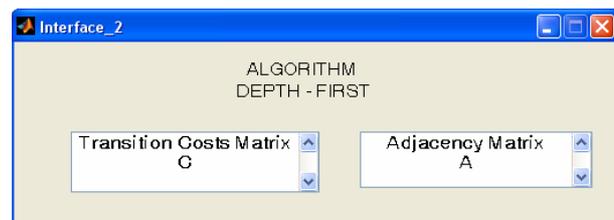


Figure 6: The Input Date of Algorithm Depth-First

As an example, the transitions costs matrix is:

$$C = \begin{bmatrix} 5 & 10 & 10 & 20 & 30 & 14 & 15 & 10 & 17 & 30 \\ 15 & 4 & 17 & 10 & 30 & 40 & 50 & 20 & 30 & 25 \\ 21 & 23 & 35 & 20 & 24 & 60 & 30 & 10 & 25 & 30 \\ 41 & 20 & 35 & 36 & 20 & 42 & 10 & 46 & 9 & 8 \\ 16 & 18 & 11 & 65 & 10 & 14 & 25 & 65 & 40 & 70 \\ 14 & 25 & 32 & 51 & 20 & 30 & 65 & 21 & 41 & 20 \\ 36 & 25 & 14 & 15 & 34 & 26 & 10 & 32 & 19 & 18 \\ 30 & 25 & 26 & 81 & 24 & 65 & 61 & 30 & 27 & 38 \\ 14 & 25 & 34 & 51 & 20 & 42 & 13 & 51 & 62 & 80 \\ 5 & 12 & 43 & 62 & 30 & 20 & 15 & 81 & 49 & 30 \end{bmatrix}$$

The adjacency matrix (A) is the same as the one considered for Chen algorithm (15). The software application determines all the possible Hamiltonian paths and computes then the minimum costs one (Fig. 7).

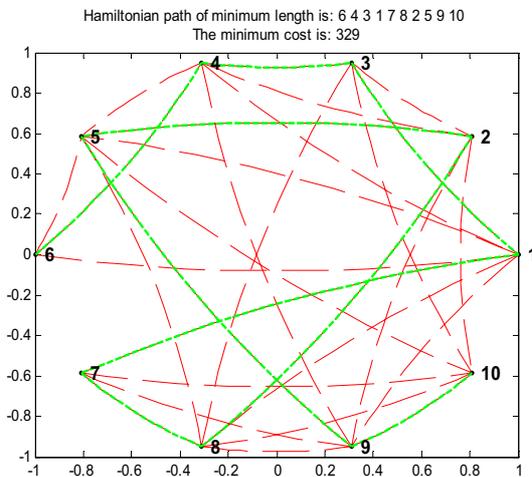


Figure 7: Minimum Costs Hamiltonian Path

If no Hamiltonian path was found, the application displays “No Hamiltonian Path” (Fig.8).

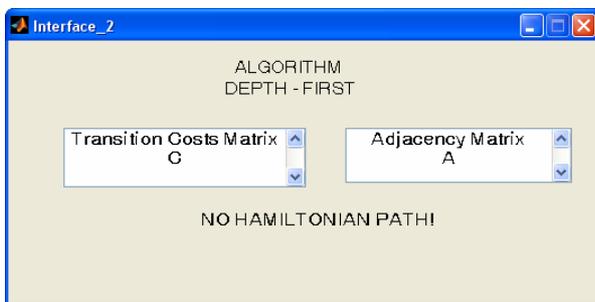


Figure 8: The Message “NO HAMILTONIAN PATH”

If another transitions costs matrix is considered,

$$C_1 = \begin{bmatrix} 10 & 5 & 15 & 20 & 30 & 14 & 15 & 10 & 17 & 30 \\ 15 & 4 & 17 & 10 & 30 & 40 & 70 & 25 & 30 & 25 \\ 21 & 23 & 35 & 30 & 24 & 20 & 30 & 10 & 25 & 30 \\ 41 & 20 & 35 & 36 & 20 & 42 & 10 & 46 & 9 & 8 \\ 16 & 50 & 11 & 10 & 10 & 14 & 25 & 65 & 40 & 70 \\ 14 & 10 & 32 & 9 & 20 & 30 & 5 & 21 & 41 & 20 \\ 36 & 5 & 14 & 5 & 10 & 45 & 10 & 32 & 19 & 18 \\ 30 & 5 & 26 & 10 & 30 & 5 & 10 & 30 & 27 & 38 \\ 14 & 25 & 34 & 51 & 20 & 15 & 13 & 51 & 4 & 80 \\ 5 & 12 & 43 & 3 & 30 & 20 & 15 & 81 & 49 & 30 \end{bmatrix}$$

The result will be the one depicted in Fig. 9.

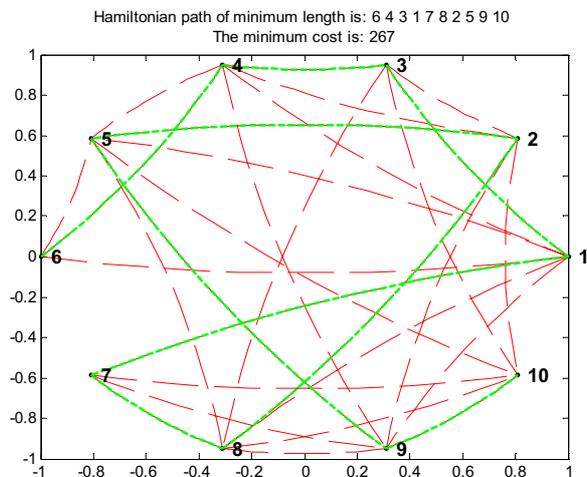


Figure 9: Minimum Costs Hamiltonian Path

As was expected, the Hamiltonian path is the same, but the costs are smaller if the matrix C_1 is considered.

CONCLUSIONS

The optimal Hamiltonian paths were determined for several applications by using algorithms known in the literature. Chen algorithm can be used only for oriented graphs which does not contain circuits. The returned result was being the single Hamiltonian path of a given graph. Depth-First Algorithm can analyze oriented, non-oriented, with or without circuits graphs. It returns the optimum Hamiltonian path from the point of view of costs. Resulted cost can even be reduced by restrictions on the transition costs matrix.

The paper presented more examples for illustration of the flexibility and the performances of the software application.

REFERENCES

- Abrudan I., 1996. “Sisteme flexibile de fabricație”, Editura Dacia, Cluj-Napoca, Romania.
- Brojboiu, M., Ivanov, V. 2003. “Ingineria sistemelor industriale”, Tipografia Universității din Craiova.

- Chakraborty, S., 2011. "Applications of the MOORA method for decision making in manufacturing environment", DOI 10.1007/s00170-010-2972-0
- Chin, F., Lam, J., Chen, I.N., 1982. "Efficient parallel algorithms for some graph problems", *Magazine Communications of the ACM*, Volume 25 Issue 9, Sept 1982, pp. 659-665.
- Fiedler, M., 1973. "Algebraic connectivity of graphs", *Czechoslovak Mathematical Journal*, Vol. 23 , No. 2, pp.298-305.
- Groza, B., 2008. "Introducere in Inteligenta Artificiala - Aplicatii cu Strategii de Cautare Neinformate si Informate", Editura Politehnica, Timisoara, Romania.
- Ivanov, V., Brojboiu, M. 2008, "Optimization of the industrial equipment emplacement by using Matlab", *Academic Journal of Manufacturing Engineering*, Vol. 6, No.1, pp.6-11.

AUTHOR BIOGRAPHIES



VIRGINIA IVANOV was born in Vela, Dolj, Romania, 1963. She was graduated in Electrical Engineering at University of Craiova, Romania, in 1986 and Doctor in Electrical Engineering in 2004. From 1986 to 1998 she worked as researcher with the Researching Institute for Motors, Transformers and Electric Equipment Craiova. In 1998 she joined the Faculty for

Electrical Engineering, Department of Electrical Equipment and technologies. She is concerned with research activities in monitoring and modeling of electrical equipments.



MARIA BROJBOIU is currently working as Professor at the University of Craiova, Electrical Engineering Faculty, Department of Electrical Energetic and Aerospace Engineering. Before that, she worked as design engineer at the Electroputere holding the Research and Development Center. She is Doctor in Science Technique – Electrical Engineering. She teaches the courses Electrical Equipment, Electrotechnologies and Industrial Systems Engineering. She published 5 books and 92 scientifically papers for different national and international conferences and symposiums. She attended other international conferences (London, Lodz – Poland, L Aquila – Italy, Nis – Yugoslavia, Ischia – Italy, Barcelona – Spania, Koblenz – Germania, Crakovia – Polonia). Concerning the managerial activities, she has been elected Dean in 2000 -2004 and Academic Secretary 2008 -2012 of the Electrical Engineering Faculty.

CURRENT CONTROL OF A VSI-FED INDUCTION MACHINE BY PREDICTIVE TECHNIQUE

Sergiu Ivanov

Vladimir Răsvan
Eugen Bobașu
Dan Popescu
Florin Stîngă

University of Craiova

Faculty of Electrical Engineering Faculty of Automation, Computer and Electronics
107 Decebal Blv., 200440, Craiova, Romania

E-mail: sergiu.ivanov@ie.ucv.ro

E-mail: [vrasvan, ebobasu, dpopescu,
florin]@automation.ucv.ro

KEYWORDS

Voltage inverter, predictive control, induction motor.

ABSTRACT

The paper deals with a technique which uses the predictive concepts in order to obtain the pulse width modulation strategy of a voltage fed inverter. After the technique is briefly described, it is applied for the case when the inverter supplies an induction motor, the reference values of the currents being obtained from a classical vector control scheme. The described technique is then simulated and the waveforms are compared with ones obtained with preset currents (bang-bang) pulse width modulation, as the behaviour of the two strategies are similar. Finally, the results are cross analysed and further actions are proposed for the work continuation.

INTRODUCTION

Nowadays, the hardware topologies of the inverters designed to supply the induction motors in variable speed drives are well crystallized. Besides quite special architectures, basically, there are the two level inverter (the classical three phase, six switches, bridge) and the multilevel inverters designed for very high power or voltage applications (Moller 2006, Steimel 2010).

Concerning control techniques of the inverters, the strategies are practically unlimited, the literature being quite rich and dynamic (Hartani 2010, Ursaru 2009, Milicevic 2013). Even the classification of these strategies can be performed by considering different criterions. We will take into account here only the source type which the inverter has: voltage or current. We speak all the time about the inverters which are supplied by a DC link having voltage source character, not real current source inverter, for which the DC link has a current source character.

We must see the modulation strategy only as a vector for obtaining the control of the whole driving system, which finally means the control of the developed torque.

Even from the basics of the vector stated by Leonhard, Blaschke and their followers in the 1970s, in rotating

references, solidar with the rotor flux, stator flux or magnetizing flux respectively, there is an obvious decoupling between the two components of the stator *current*: while the direct component acts on the flux modulus only and produces the reactive component, the quadrature component generates the torque, being the active component. The two components of the stator current must be thus controlled independently and the flux and torque generation are thus decoupled, similarly to the DC motor.

This means that, as the torque is controlled by the current components, a current source inverter is more suited for the control of the torque developed by the drive. The previous work of the authors emphasized that the field oriented control (FOC) schemes based on current source inverters (preset currents, or bang-bang modulation) are more robust to the parameters variations and have very good dynamics. The main disadvantages of this simple modulation strategy are related to the very high necessary switching frequency (available only in the low range of power), variable switching frequency (difficult to estimate the losses) and interphases dependency. Different techniques were developed for improving the strategy (sinusoidal hysteresis, multilevel hysteresis comparators, Mohseni 2010), but the variable switching frequency rests always as an disadvantage.

The technique we propose has the behaviour of preset currents inverter, but it performs the pulse width modulation with fixed frequency, which is in fact the sampling frequency of the system.

From this point of view (fixed switching frequency given by the sampling one), the proposed technique has a similarity with another very simple method for the torque control, the Direct Torque Control (DTC), suited for electrical traction applications (Takahashi and Noguchi 1986, Baader et al. 1992, Ehsani et al. 1997, Faiz et al. 1999, Haddoun et al. 2007, Ivanov 2009, Ivanov 2010). As will be shown, as the direct controlled variables are the stator currents, the behaviour of the proposed technique is much better.

The proposed technique is feasible due to the increased computational capabilities of the existing DSP which allow the implementation of the predictive control at the

level of the converters which induce the hybrid character of the overall control system of the drive. We infer that predictive control has established itself in the last 5-7 years as a very proficient form of controlling highly nonlinear and uncertain systems; moreover the most recent results show its applicability to fast processes among which drives and their converters have a central position (Seo et al. 2009, Prieur and Tarbouriech 2011, Geyer et al. 2008, Mariethoz et al. 2010, Geyer et al. 2009, Trabelsi et al. 2008, Shi et al. 2007, Rodriguez et al. 2007, Larrinaga et al. 2007, Richter et al. 2010, Almer et al. 2010).

The paper will briefly present in the first section the basics of the predictive control. The principle of the predictive control applied to a three phase inverter will be presented in Section 2. Section 3 will detail the predictive control of the inverter fed induction machine drive. Section 4 will present the results of the simulations performed based on a Simulink model. The results will be compared with the ones obtained with the bang-bang strategy and with DTC. Finally, conclusions will be issued and ideas for continuation will be pointed out.

BASICS OF PREDICTIVE CONTROL

The model predictive control is a control technique which has been successfully implemented in industry. The predictive control techniques were used to control both continuous as well as discrete systems (Camacho and Bordons 2004, Bemporad 2007, Lazăr 2006, Maciejowski 2000, Stinga 2012).

The predictive control is derived from optimal control, yet, in this case the optimal control problem involves additional constraints.

The predictive control techniques require solving an open loop optimal control problem, taking into account constraints on input, state and/or output variables. At every moment k , the measured variables and the model of the process are used to compute (to predict) the future behaviour of the system over a prediction horizon N (Fig. 1).

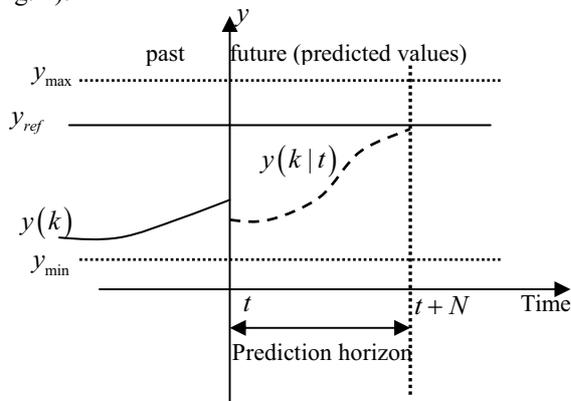


Figure 1: Evolution of System Output Using Predictive Control Strategy (Stinga 2012)

This task is accomplished by determining a set of future control inputs such that the objectives and the system constraints are satisfied. The control input is determined by minimization of a cost function over a time horizon N_c .

Generally, the cost function used in predictive control is defined as follows:

$$J(k) = \sum_{t=1}^N \left\| y(k|t) - y_{ref}(k) \right\|_{Q(t)}^2 + \sum_{t=1}^{N_c} \left\| u(k|t) \right\|_{R(t)}^2, \quad (1)$$

subject to constraints specified on the inputs, outputs and input increments (Fig. 2):

$$\begin{aligned} u_{\min} &\leq u(k) \leq u_{\max}, \\ y_{\min} &\leq y(k) \leq y_{\max}, \end{aligned}$$

where:

- $Q(t)$ - positive definite error weighting matrix;
- $R(t)$ - positive semi-definite control weighting matrix;
- $y(k|t)$ - vector of predicted output signals;
- $y_{ref}(k)$ - vector of future set points;
- $u(k|t)$ - vector of future control inputs;
- N - prediction horizon;
- N_c - control horizon.

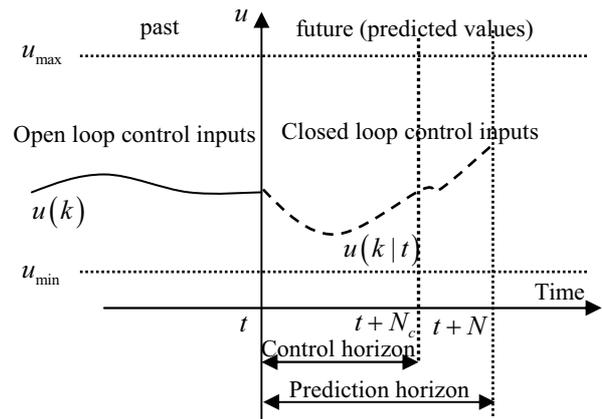


Figure 2: The Control Inputs Applied to the System Using the Predictive Control Strategy (Stinga 2012)

PREDICTIVE CONTROL OF THE THREE PHASE INVERTER

The basic ideas of the predictive control of the three phase bridge inverter are presented in Rodriguez 2012, for a simple R-L load.

The predictive command of the inverter is facilitated by the limited number of possible future states. In fact, the inverter can have only eight different topologies (Fig. 3). These eight different topologies determine seven positions of the voltage phasor (Fig. 4). It is to note that two topologies (7 and 8) are equivalent and determine the same position of the voltage phasor. In practice, one of the two is chosen depending on the actual state of the inverter in order to minimize the number of switches. If the actual state is one of 2, 4 or 6 and the zero phasor must be obtained, the topology 7 will be chosen.

Contrary, if the actual state is one of 1, 3 or 5 and the zero phasor must be obtained, the topology 8 will be chosen.

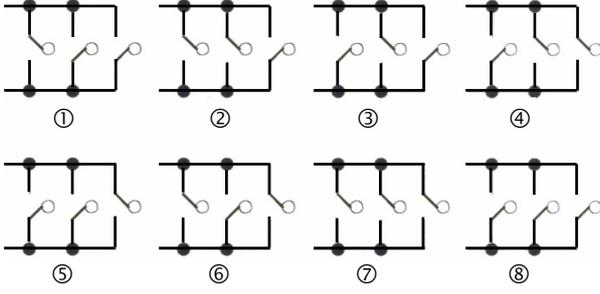


Figure 3: Possible Topologies of the Three Phase Bridge

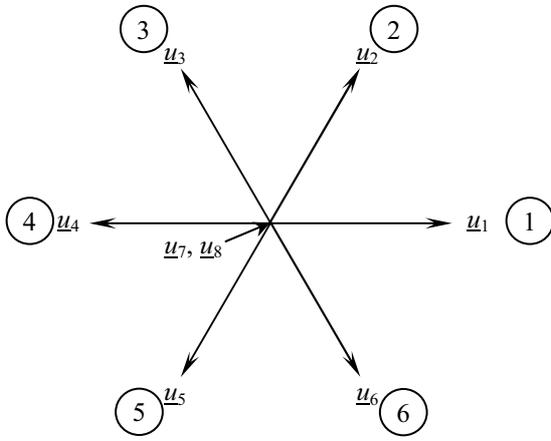


Figure 4: Positions of the Voltage Phasor

The operating principle is to compute at each sampling step the estimations of the (α, β) components of the currents, $i_{\alpha}^e, i_{\beta}^e$, for all the seven different values of the voltages corresponding to different topologies. Then, the topology that will be applied for the next sampling period will be chosen the one which minimises the cost function

$$J(k) = \left| i_{\alpha}^*(k+1) - i_{\alpha}^e(k+1) \right| + \left| i_{\beta}^*(k+1) - i_{\beta}^e(k+1) \right|, \quad (2)$$

where $i_{\alpha}^*, i_{\beta}^*$ are the preset values of the (α, β) components of the currents.

PREDICTIVE CONTROL OF THE VSI FED INDUCTION MOTOR DRIVE

The principle described above must be applied considering as load of the inverter, an induction machine.

At each sampling period, having as initial conditions the actual values of the stator and rotor currents components, $i_{s\alpha}, i_{s\beta}, i_{r\alpha}, i_{r\beta}$, the state equation model of the motor (3) is integrated for all the seven different

values of the input vector, which consists in the voltage components $[uu] = [u_{s\alpha}, u_{s\beta}, u_{r\alpha}, u_{r\beta}]^T$,

$$\frac{d}{dt}[i] = [ML]^{-1} ([uu] - ([MR] + [MXr]) \cdot [i]), \quad (3)$$

where:

$[i] = [i_{s\alpha}, i_{s\beta}, i_{r\alpha}, i_{r\beta}]^T$ - the stator and rotor (α, β) current components;

$$[ML] = \begin{bmatrix} L_s & 0 & L_m & 0 \\ 0 & L_s & 0 & L_m \\ L_m & 0 & L_r & 0 \\ 0 & L_m & 0 & L_r \end{bmatrix} - \text{the inductances matrix,}$$

whose components are:

L_s, L_r - total stator and rotor inductances;

L_m - mutual inductance;

$$[MR] = \begin{bmatrix} R_s & 0 & 0 & 0 \\ 0 & R_s & 0 & 0 \\ 0 & 0 & R_r & 0 \\ 0 & 0 & 0 & R_r \end{bmatrix} - \text{the resistances matrix;}$$

$$[MXr] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & P\omega_r L_m & 0 & P\omega_r L_r \\ -P\omega_r L_m & 0 & -P\omega_r L_r & 0 \end{bmatrix} - \text{the reactance matrix, whose terms depend on the}$$

P - number of pairs of poles and

Ω_r - mechanical speed of the rotor.

The model (3) is completed with the movement equation which must be also be integrated at each sampling period

$$\frac{d\omega_r}{dt} = \frac{1}{J} \left[\frac{3}{2} PL_m (i_{s\beta} i_{r\alpha} - i_{s\alpha} i_{r\beta}) - T_s \right], \quad (4)$$

where:

J - total inertia at the motor shaft;

T_s - static torque applied to the rotor shaft.

It result seven sets of state variables estimations

$$[i_{s\alpha}^e, i_{s\beta}^e, i_{r\alpha}^e, i_{r\beta}^e, \omega_r^e]_{1..7}$$

and the cost function (2) is computed for the stator currents components

$$J(k)|_{1..7} = |i_{s\alpha}^* - i_{s\alpha}^e| + |i_{s\beta}^* - i_{s\beta}^e|. \quad (5)$$

The next topology of the inverter is chosen the one which corresponds to the minimum of the seven values given by (5).

At this stage, there are not considered the both topologies which determine the zero voltage phasor (7 and 8), because the aim is only to determine the next stator voltages which minimises the cost function, not the real topology of the inverter. This presents only practical importance at the implementation stage.

The preset values of the stator currents components $i_{s\alpha}^*$, $i_{s\beta}^*$ are the results of a classical FOC of the induction machine supplied by a preset currents inverter, Fig.5.

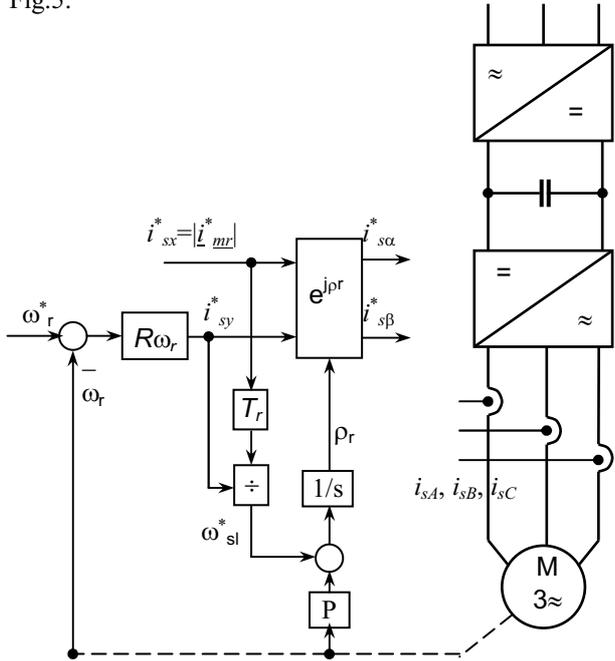


Figure 5: The Classical FOC of the Induction Machine Supplied by Preset Currents Inverter

The block which estimates the currents and computes the cost function (5) must be placed between the rotation transformation block e^{jpr} and the inverter.

SIMULINK MODEL OF THE DRIVE

The complete Simulink model of the drive is depicted in Fig. 6.

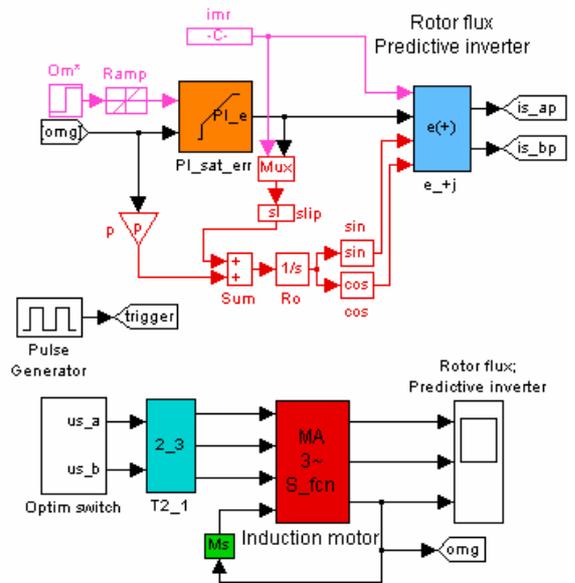


Figure 6: Simulink Model of the Drive

As can be seen, the classical FOC (upper part of the figure) outputs the preset values of the two stator currents components. They are applied (by the way of GoTo tags) to the *Optim switch* block. This block (Fig. 7) computes all the currents in the model (3) and the speed (4), for all the seven possible values of the stator voltages (blocks I_1 to I_7).

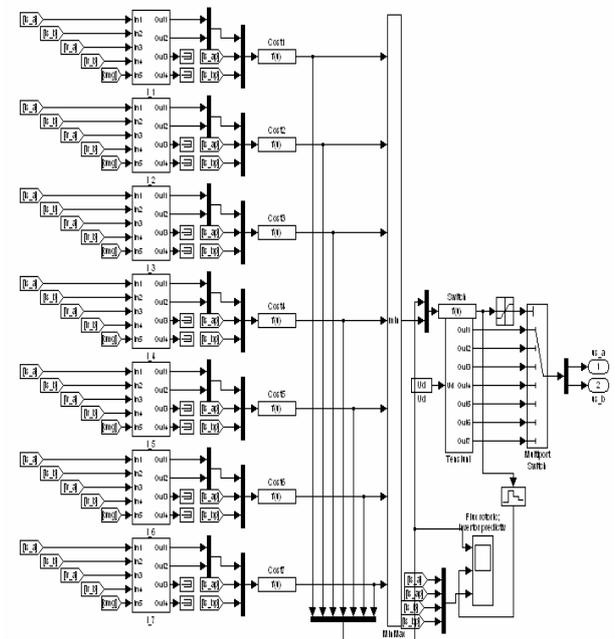


Figure 7: The Optim Switch Block

Each of the seven blocks computes the estimated values of the stator and rotor currents components based on (3), the integrators being reset with the actual values of the four currents components (Fig. 8).

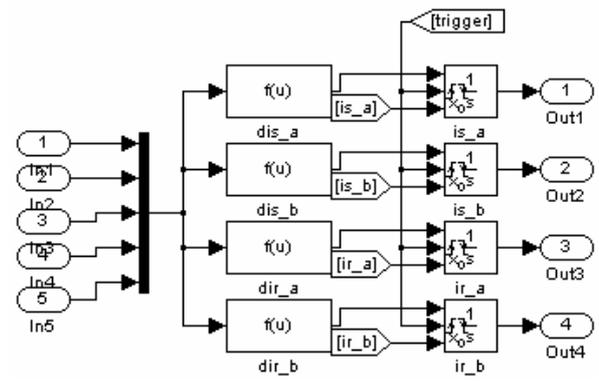


Figure 8: The Computing of the Stator and Rotor Currents Components

The cost function (5) is then computed for all the seven possible values of the stator voltages, the minimum of the seven is determined and it is identified the topology which determines that this minimum is achieved. The output vector $u_{s\alpha}$, $u_{s\beta}$, the rotor being considered squirrel cage and consequently, $u_{r\alpha} = u_{r\beta} = 0$.

SIMULATION RESULTS AND COMPARISONS

The simulation results of three types of command are presented in Fig. 9, 10 and 11, in all cases the simulation step being constant and equal to $100 \mu\text{s}$. Only the phase currents are plotted as results of the simulations, the comparison being performed from this point of view. Of course, a better (smaller) ripple of the phase currents determines better overall behaviour of the drive (smaller torque ripple, greater average torque and better dynamics).

Fig. 9 plots a detail of the currents obtained by using the presented technique.

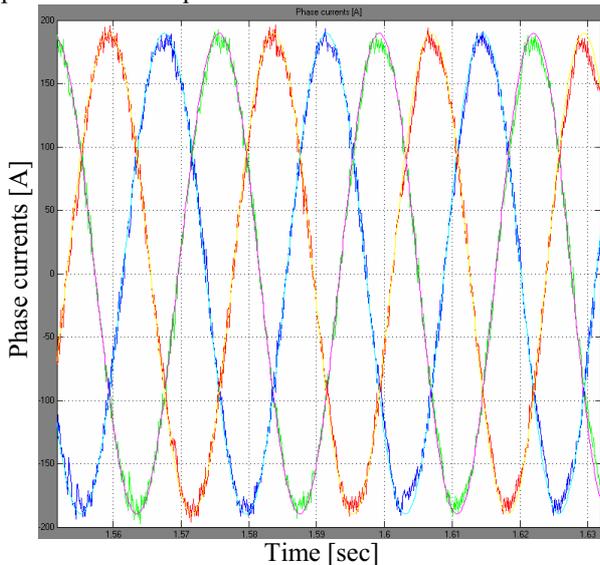


Figure 9: Stator Currents for Predictive Command of the Inverter

These waveforms must be compared with the ones obtained with a classical bang-bang modulator (preset currents), but with fixed switching frequency (the same as for predictive control). In this case (Fig. 10), the ripple of the currents is significantly higher.

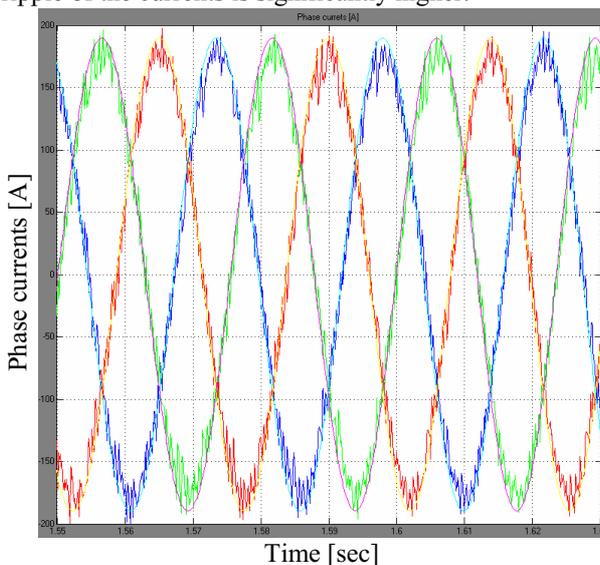


Figure 10: Stator Currents for Preset Currents Modulation

We notice that, for the same constant switching frequency, the current ripple can arise to be four times larger (40 A, compared with 10A). This is because, for preset currents (bang-bang) modulation, the switches are obtained independently on the three phases. For the predictive modulation, the topology of the inverter is chosen globally, as the one which minimizes the currents errors.

Finally, the waveforms are compared with the ones resulted when a classical DTC controls the induction motor. Once again, the sampling period is the same $100 \mu\text{s}$. We make this comparison due to the similarity of the commands: the both determine the next stator voltage phasor, but considering different criteria.

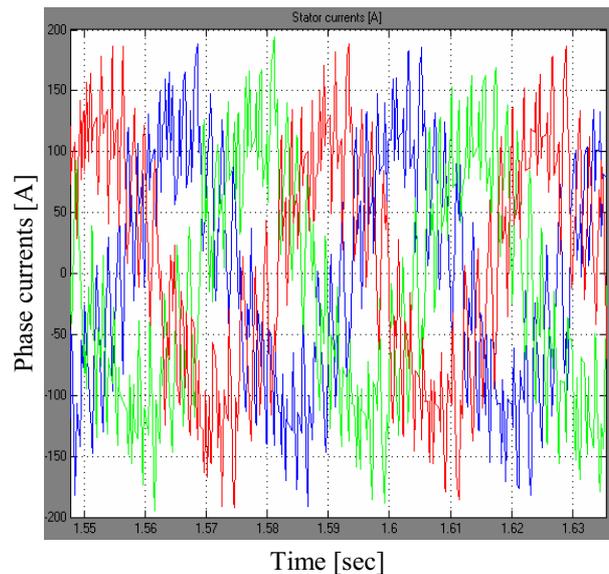


Figure 11: Stator Currents for Classical DTC

We notice that the results are the worst, from the point of view of currents ripple.

All the simulations were performed for a motor with the rated values:

- power: 55 kW;
- speed: 981 r/min;
- L-L voltage: 380 V;
- stator phase resistance: 0.068Ω ;
- rotor phase resistance: 0.044Ω ;
- stator leakage inductance: 0.5 mH;
- rotor leakage inductance: 0.5 mH;
- magnetising inductance: 21.5 mH;
- pairs of poles: 3.

CONCLUSIONS

The paper describes a strategy for pulse width modulation of the inverters which supply induction machines, based on the predictive technique.

A Simulink model of the drive is presented and the results of the simulation are compared, for similar conditions (same fixed step), with other two modulation techniques: preset currents and DTC. The currents ripple

is the smallest when the proposed modulation technique is used. The consequences are favourable in what concern the torque ripple and general dynamic behaviour. The goal of the research is to implement the predictive control for an industrial drive which will be offered on the market.

Further research will be focused on optimizing the proposed technique, in order to reduce the computing time.

ACKNOWLEDGMENTS

Authors wish to thank the UEFISCDI and their partners in the HYDICO project (PN-II-PT-PCCA-2011-3.2-1082), in the frame of which their study has been performed.

REFERENCES

- Almer S. et al. 2010. "Piecewise Affine Modeling and Control of a Step-Up DC-DC Converter". *American Control Conference Paper ThB05.3*.
- Baader U., Depenbrock M., Gierse G. 1992. "Direct self control (DSC) of inverter-fed induction machine: a basis for speed control without speed measurement". In *IEEE Transactions on Industrial Applications*, vol. 28, pp. 581–588.
- Ehsani M. et al. 1997. "Propulsion system design of electric and hybrid vehicles". In *IEEE Trans. Industrial Electronics*, vol. 45, nr.1, pp 19-27.
- Faiz J. et al. 1999. "Direct torque control of induction motor for electric propulsion systems," In *International Journal on Power Systems*, vol. 51, pp. 95–101.
- Geyer T. et al. 2008. "Hybrid Model Predictive Control of the Step Down DC-DC Converter". In *IEEE Trans. Contr. Syst. Technol.* Vol. 16, no.6, pp. 1112-1124.
- Geyer T. et al. 2009. "Model Predictive Direct Torque Control—Part I: Concept, Algorithm, and Analysis". In *IEEE Transactions on Industrial Electronics*. Vol. 56, no.6, pp. 1894-1905.
- Haddoun A., Benbouzid M., Dialo D., Abdessemed R., Ghouili J., Srairi K. 2007. "A loss-minimization DTC Scheme for EV Induction Motor" In *IEEE Transactions on vehicle technology*, vol.56, nr.1, pp.81-88.
- Hartani, K.; Miloud, Y. 2010. "Control Strategy for Three Phase Voltage Source PWM Rectifier Based on the Space Vector Modulation" In *Advances in Electrical and Computer Engineering*, vol. 10, no. 3, pp. 61-65, 2010, doi:10.4316/AECE.2010.03010
- Ivanov, S. 2009. "The influence of the sampling period on the performance of the regulation by DTC of the induction motor". In *Proceedings of the 23rd European Conference on Modelling and Simulation*, Madrid, Spain, 776-780.
- Ivanov, S., 2010. "Continuous DTC of the Induction Motor". In *Advances in Electrical and Computer Engineering*, vol. 10, no. 4, 149-154.
- Ivanov, S.; V. Răsvan; E. Bobașu; D. Popescu; F. Stîngă; V. Ivanov. 2013. "Predictive versus vector control of the induction motor". In *Proceedings of the 27th European Conference on Modelling and Simulation (27-31.05.2013, Ålesund, Norway)*, ISBN: 978-0-9564944-6-7 / ISBN: 978-0-9564944-7-4 (CD), 434-440.
- Larrinaga S.A. et al. 2007. "Predictive Control Strategy for DC/AC Converters Based on Direct Power Control". In *IEEE Trans. Ind. Electronics*. Vol. 54, no.3, pp. 1261-1270.
- Mariethoz S. et al. 2010. "Comparison of Hybrid Control Techniques for Buck and Boost DC-DC Converters". In *IEEE Trans. Contr. Syst. Technol.* Vol. 18, no.5, pp. 1126-1145.
- Milicevic, D.; Katic, V.; Corba, Z.; Greconici, M. 2013. "New Space Vector Selection Scheme for VSI Supplied Dual Three-Phase Induction Machine". In *Advances in Electrical and Computer Engineering*, vol. 13, no. 1, pp. 59-64, 2013, doi:10.4316/AECE.2013.01010.
- Mohseni, M.; Islam, S.M. 2010. "A New Vector-Based Hysteresis Current Control Scheme for Three-Phase PWM Voltage-Source Inverters". In *IEEE Transactions on Power Electronics*, Volume 25 , Issue: 9, Sept. 2010, 2299 – 2309.
- Moller, D., Schlegel, C., Velaro, "Further Development of the ICE for Worldwide Use" *Elektrische Bahnen 104* (2006), Nr. 5, pp. 258-263.
- Prieur C. and Tarbouriech S. 2011. "New directions in hybrid control systems" (editorial). In *Int. Journal Robust Nonlin. Control*. Vol. 21, pp. 1063-1065.
- Richter S. et al. 2010. "High-Speed Online MPC Based on a Fast Gradient Method Applied to Power Converter Control". *American Control Conference Paper FrA01.6*.
- Rodriguez J. et al. 2007. "Predictive Current Control of a Voltage Source Inverter". In *IEEE Trans. Ind. Electronics*. Vol. 54, no.1, pp. 495-503.
- Rodriguez J.; Cortes, P. 2012. *Predictive Control of Power Converters and Electrical Drives*. Wiley.
- Seo S. et al. 2009. "Hybrid Control System for Managing Voltage and Reactive Power". In *the JEJU Power System, Journal of Electrical Eng. and Technol.* Vol. 4, no.4 pp. 429-437.
- Shi X.L. et al. 2007. "Implementation of Hybrid Control for Motor Drives". In *IEEE Trans. Ind. Electronics*. Vol. 54, no. 4, pp. 1946-1952.
- Steimel A. 2010. "Power-Electronics Issues of Modern Electric Railway Systems". In *Advances in Electrical and Computer Engineering*, vol. 10, no. 2, pp. 3-10, doi:10.4316/AECE.2010.02001.
- Takahashi I. and Noguchi T. 1986. "A new quick-response and high efficiency control strategy of an induction motor". In *IEEE Transactions on Industrial Applications*. vol. IA-22, no.5, pp. 820-827.
- Trabelsi M. et al. 2008. "Hybrid Control of a Three-Cell Converter Associated to an Inductive Load". *IEEE paper 978-1-4244-1668-4/08*.
- Ursaru, O.; Aghion, C.; Lucanu, M.; Tigaeru, L. 2009. "Pulse width Modulation Command Systems Used for the Optimization of Three Phase Inverters". In *Advances in Electrical and Computer Engineering*, vol. 9, no. 1, pp. 22-27, 2009, doi:10.4316/AECE.2009.01004.
- Camacho E.F. and Bordons C. 2004. *Model predictive control*. Springer-Verlag.
- Bemporad A. 2007. *Model predictive control of hybrid systems*, 2nd HYCON PhD. School on Hybrid Systems, Siena.
- Lazăr M. 2006. *Model Predictive Control of Hybrid Systems: Stability and Robustness*, Ph.D. Thesis, Eindhoven, Holland.
- Maciejowski J.M. 2000. *Predictive control with constraints*, Prentice Hall.
- Stinga F. 2012. *Control strategies for hybrid systems. Applications*, Ph.D. Thesis, Craiova, Romania.

AUTHOR BIOGRAPHIES



SERGIU IVANOV was born in Hunedoara, Romania and went to the University of Craiova, where he studied electrical engineering. He obtained his degree in 1986. He worked for the Institute for Research in Motors, Transformers and Electrical Equipment Craiova before moving in 1991 to the University of Craiova. He obtained his PhD in 1998 with a topic in the field of the control of the electric drives systems. He is involved in modelling of the electromechanical systems.



VLADIMIR RASVAN graduated from the Polytechnic Institute of Bucharest, Romania in 1967 (Automatic Control) and obtained his Ph.D. in System Theory in 1972. After a 10 years career in applied research for control in Power systems, he became an associate professor (1982) and eventually professor (1990) at the University of Craiova. His main scientific interests are in mathematical approaches for dynamics in engineering systems. He is author of 7 books and some 200 papers published in scientific journals and proceedings of scientific/technical conferences.



EUGEN BOBAȘU received the B.S. and M.Sc. degrees (1977), both in automatic control, and the Ph.D. degree in control systems (1997) from the University of Craiova, Romania. Since 1981 he is with

the University of Craiova, where he is currently Professor in the Department of Automatic Control.

He is involved in national and international research projects in the field of modelling, identification and hydraulics. His present research interests are on modelling of complex systems and identification of nonlinear systems. He has published more than 90 journal and conference papers, and he is author or co-author of 5 books.

Prof. Bobașu is member of IEEE, SRAIT and of ARR.



DAN POPESCU received the B.S. and M.Sc. degrees (1977), both in automatic control, and the Ph.D. degree in control systems (1997) from the University of Craiova, Romania. Since 1981 he is with the University of Craiova, where he is currently Professor in the Department of Automation, Electronics and Mechatronics. His present research interests are on robust control, time delay systems and predictive control. Prof. POPESCU is member of IEEE and IFAC TC 2.5 “Robust Control”.



FLORIN STINGA was born in Craiova, Romania. He received the B. Eng., M.S. and Ph.D. degrees in system engineering, all from University of Craiova, in 2000, 2003 and 2012. Currently, he is Assistant Professor in the Department of Automation, Electronics and Mechatronics at the Faculty of Automation, Computers and Electronics, Craiova. His researches interested are in hybrid dynamical systems and embedded systems.

MODELING AND CONTROL OF BERRY PHASE IN QUANTUM DOTS

Sanjay Prabhakar¹, Roderick Melnik¹ and Ali Sebetci²

¹M²NeT laboratory, Wilfrid Laurier University, Waterloo, ON, Canada

²Department of Mechanical Engineering, Mevlana University, Konya, Turkey

KEYWORDS

Semiconductor quantum dots, Finite Element Method (FEM), Berry phase, spin-orbit coupling, quantum mechanical transport.

ABSTRACT

We study numerically the Berry phase in semiconductor quantum dots (QDs) that is induced by letting the dots to move adiabatically in a closed loop in the 2D plane along the circular trajectory. We show that the Berry phase is highly sensitive to the Rashba and Dresselhaus spin-orbit lengths. Based on the Finite Element Method, we solve the Schrödinger equation and investigate the evolution of the spin dynamics during the adiabatic transport of the QDs in the 2D plane along circular trajectory. Results of numerical simulations are discussed in detail, indicating that this work might be used for the realization of solid state quantum information processing.

INTRODUCTION

Manipulating the single electron spins in QDs through the non-Abelian geometric phases has attracted considerable attention since the pioneering work of Berry (Aleiner & Fal'ko 2001, Wang & Zhu 2008, Yang & Hwang 2006, Eric Yang 2006, Yang 2007, Berry 1984). For a system of degenerate quantum states, Wilczek and Zee showed that the geometric phase factor is replaced by a non-Abelian time dependent unitary operator acting on the initial states within the subspace of degeneracy (Wilczek & Zee 1984, Prabhakar et al. 2010). Since then the geometric phase has been measured experimentally for a variety of systems such as quantum states driven by a microwave field (Pechal et al. 2012) and qubits with tilted magnetic fields (Berger et al. 2012, Leek et al. 2007). Manipulation of the spin qubits through the Berry phase implies that the injected data can be read out with different phase that can be topologically protected from the outside world (Das Sarma et al. 2005, Hu & Das Sarma 2000, Loss et al. 1990, Tserkovnyak & Loss 2011, San-Jose et al. 2008). Several recent reviews of the Berry phase have been presented in Refs. (Xiao et al. 2010, Nayak et al. 2008). One of the promising research proposals for building a solid state topological quantum computer is that the accumulated Berry phase in QD system can be manipulated with the interplay between the Rashba-Dresselhaus spin-orbit couplings (San-Jose et al. 2008, Aleiner & Fal'ko 2001). The Rashba spin-orbit coupling arises from the asymmetric triangular quantum well along the growth direction and the Dresselhaus spin-orbit coupling arises due to bulk inversion asymmetry in the crystal lattice (Bychkov & Rashba 1984, Dresselhaus 1955). A

recent work by Bason et al. shows that the Berry phase can be measured for a two level quantum system in a superadiabatic basis comprising the Bose-Einstein condensates in optical lattices (Bason et al. 2012).

The geometric phase induced on the wavefunctions of quantum states during the adiabatic movement of the physical system plays an important role in numerous quantum computing and quantum information processing. When the state vector of a quantum system undergoes in a cyclic evolution and returns in its initial physical state then its wave function can acquire a geometric phase factor in addition to the familiar dynamic phase (Berry 1984, Prabhakar et al. 2010, Wang & Zhu 2008). If the cyclic change of the system is adiabatic then after one complete rotation of the physical system acquire an additional phase factor which is known as Berry phase (Berry 1984). Recently, it has also been shown that the geometric phase can be induced on the electron spin states in QDs by moving the dots adiabatically in a closed loop in the 2D plane with the application of gate controlled electric field (Prabhakar et al. 2010, San-Jose et al. 2008). Furthermore, the authors in Refs (Bednarek et al. 2012, 2008, Bednarek & Szafran 2008) have recently proposed to build a QD device in the absence of the magnetic fields that can perform the quantum gate operations (NOT gate, Hadamard gate and Phase gate) with the application of the externally applied gate potential modulated by a sinusoidal varying potential. All these problems can be studied efficiently with the tools of mathematical modeling, once an adequate physical model is constructed. In this paper, we focus on modeling of transport of the electron spin states in QDs in presence of the externally applied magnetic fields along z-direction in a closed loop in the 2D plane with the application of time dependent distortion potential. Based on our model, we investigate the interplay between the Rashba and the Dresselhaus spin-orbit lengths on the scalar Berry phase (Yang & Hwang 2006, Wu et al. 2011). The transport of the dots is carried out very slowly so that the adiabatic theorem can be applied on the evolution of the spin dynamics. We show that the Berry phase in QDs can be engineered and can be manipulated with the application of the spin-orbit couplings through gate controlled electric fields. We solve the time dependent Schrödinger equation and investigate the evolution of spin dynamics in QDs. Details of the corresponding mathematical model and computational methodology are provided.

MATHEMATICAL MODEL

The model construction starts from the two band Kane Hamiltonian of an electron in QDs in the plane of a 2 Dimensional Electron Gas (2DEG) in the presence of an external

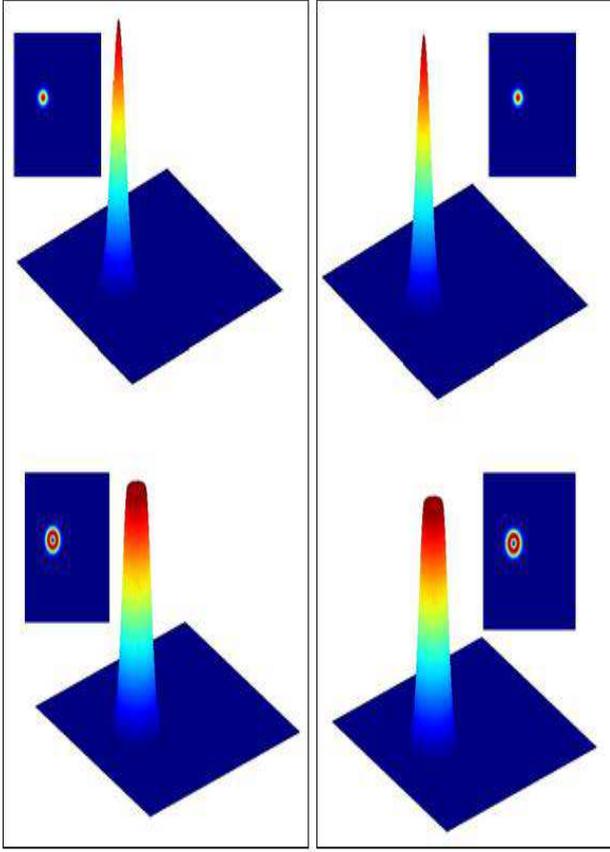


Fig. 1. Modeling results for four lowest states of the wavefunction squared in GaAs quantum dots. Here we chose $B = 1T$ and $E_0 = 5 \times 10^3 \text{ V/cm}$. Note that the spin split wavefunctions shown in the left and right columns look identical. However, their energy eigenvalues are different which can be used for the design of quantum dots with different g-factors and Berry phases (see Figs. 2 and 5).

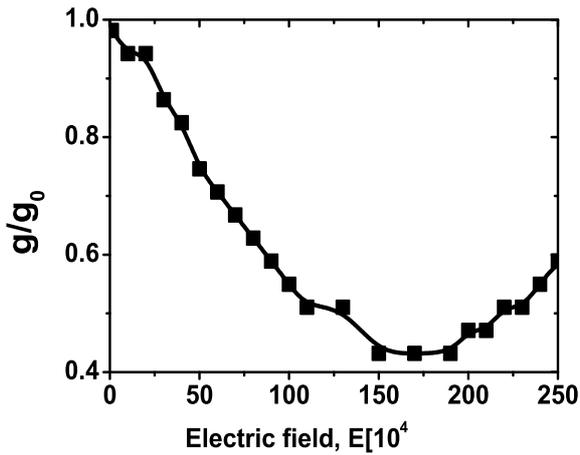


Fig. 2. g-factor (absolute value) vs applied electric fields with no time dependent distortion potential. We chose $g_0 = -0.44$, $m = 0.067$, $\gamma_R = -0.044 \text{ nm}^2$ and $\gamma_D = -0.0026 \text{ eV} \cdot \text{nm}^3$.

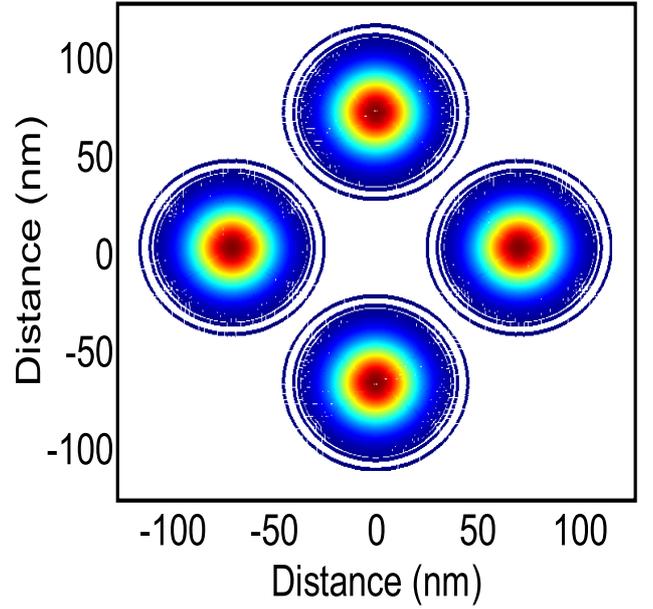


Fig. 3. Contour plots of the realistic electron wave function of GaAs QDs that are adiabatically transported along the circular trajectory under the influence of externally applied time dependent gate potential. We chose the amplitude $f_0 = 5 \times 10^3 \text{ V/cm}$, electric field $E = 10^5 \text{ V/cm}$, $B = 1T$ and QD radius $\ell_0 = 20 \text{ nm}$.

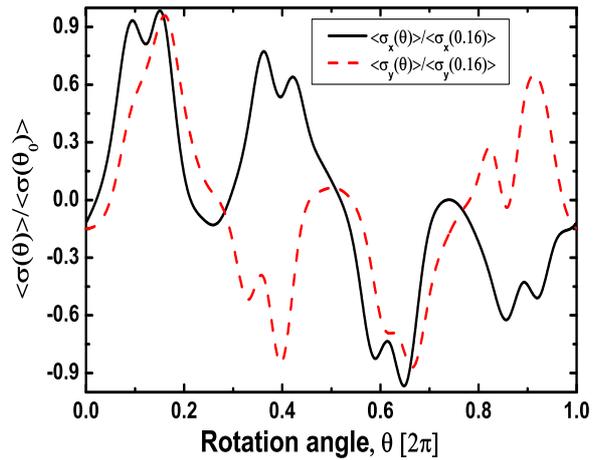


Fig. 4. Evolution of spin dynamics during the adiabatic transport of the GaAs quantum dots. We chose $E = 5 \times 10^5 \text{ V/cm}$ and the rest of the parameters are chosen the same as in Fig. 3.

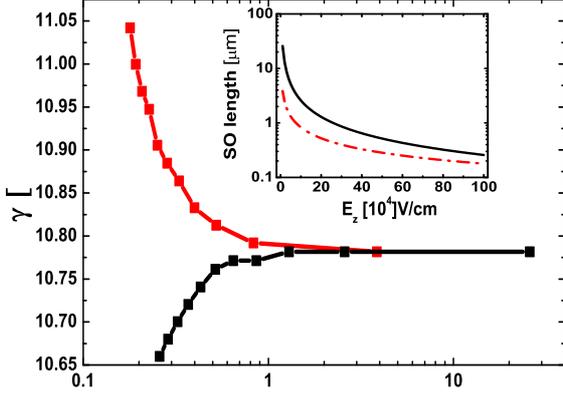


Fig. 5. Berry phase vs SO lengths on spin state $|0,0,+1\rangle$. Here we chose $\ell_0 = 20$ nm, $B = 1$ T and total enclosed adiabatic area is considered as $\pi d^2/4 = 7.85 \times 10^{-7} \text{eV}^2/\text{nm}^2$.

magnetic field B , along the z -direction in III-V semiconductor QDs can be written as (Prabhakar & Reynolds 2009, Prabhakar et al. 2011)

$$H = H_{xy} + H_R + H_D, \quad (1)$$

where the Hamiltonians H_R and H_D are associated with the Rashba and the Dresselhaus spin-orbit couplings and H_{xy} is the Hamiltonian of the electron along the lateral direction in the plane of the 2DEG. H_{xy} can be written as

$$H_{xy} = \frac{\vec{P}^2}{2m} + \frac{1}{2}m\omega_0^2(x^2 + y^2) + f(t) + \frac{\hbar}{2}\omega_z\sigma_z, \quad (2)$$

where $\vec{P} = \vec{p} + e\vec{A}$ is the kinetic momentum operator, $\vec{p} = -i\hbar(\partial_x, \partial_y, 0)$ is the canonical momentum operator and \vec{A} is the vector potential in the symmetric gauge, $\omega_z = g_0\mu_B B/\hbar$ is the Zeeman frequency and g_0 is the bulk g -factor. Here, $-e < 0$ is the electronic charge, m is the effective mass of the electron in the conduction band, μ_B is the Bohr magneton, σ_z is the Pauli spin matrix along z -direction. Also, $\omega_0 = \frac{\hbar}{m\ell_0^2}$ is a parameter characterizing the strength of the confining potential and ℓ_0 is the radius of the QD. The time dependent function $f(t)$ is the distortion potential that can be used to let the dot to move adiabatically in a closed loop in the 2D plane without disturbing the spin splitting energy difference. We use the functional form of $f(t)$ in our theoretical model as (Yang & Hwang 2006)

$$f(t) = eF_x(t)x + eF_y(t)y, \quad (3)$$

where $F_x = f_0 \cos(\omega t)$, $F_y = f_0 \sin(\omega t)$, f_0 is the amplitude and ωt varies from 0 to 2π .

The Hamiltonians associated with the Rashba-Dresselhaus spin-orbit couplings can be written as (Bychkov & Rashba 1984, Dresselhaus 1955)

$$H_R = \frac{\alpha_R}{\hbar} (\sigma_x P_y - \sigma_y P_x), \quad (4)$$

$$H_D = \frac{\alpha_D}{\hbar} (-\sigma_x P_x + \sigma_y P_y). \quad (5)$$

The strength of the Rashba-Dresselhaus spin-orbit couplings is characterized by the parameters α_R and α_D which are given by

$$\alpha_R = \gamma_R e E, \quad \alpha_D = 0.78 \gamma_D \left(\frac{2me}{\hbar^2} \right)^{2/3} E^{2/3}. \quad (6)$$

Finally we write the two coupled Schrödinger equations as:

$$-i\hbar\partial_t \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} h_{11} + \frac{\Delta}{2} & h_{12} \\ h_{21} & h_{22} - \frac{\Delta}{2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad (7)$$

where $\Delta = g_0\mu_B B/2$ and

$$h_{11} = -\frac{\hbar^2}{2m} (\partial_x^2 + \partial_y^2) + \frac{1}{2}m\Omega^2 (x^2 + y^2) + f(t) - \frac{i\hbar\omega_c}{2} (y\partial_x - x\partial_y), \quad (8)$$

$$h_{12} = \hbar\alpha_R (\partial_x - i\partial_y) + \hbar\alpha_D (i\partial_x - \partial_y). \quad (9)$$

Also, $h_{11} = h_{22}$ and h_{21} is hermitian conjugate of (h_{12}) . Finally we define the g -factor of electron in quantum dots by the expression

$$g = \frac{\varepsilon_1 - \varepsilon_2}{\mu_B B}, \quad (10)$$

where ε_1 and ε_2 are the ground and first excited states eigenvalues of the corresponding two coupled Schrödinger equations (7). In principle, one can design GaAs/AlGaAs quantum dots and vary the g -factor of the dots by letting the wavefunction of electrons to penetrate from one material (GaAs) into the other material (AlGaAs) (Prabhakar & Reynolds 2009).

We now turn to the calculation of the Berry phase in QDs. According to works of Berry, if parameters contained in the Hamiltonian of a quantal system are adiabatically carried around a closed loop, an extra geometric phase (Berry phase) is induced in addition to the familiar dynamical phase (Berry 1984, Prabhakar et al. 2010). A slow variation of such parameters along a closed path C will return the system to its original energy eigenstate with an additional phase factor $\exp\{i\gamma_n(C)\}$. More specifically, the state acquires phases after a period of the cycle T as

$$|\Psi_n(T)\rangle = \exp\left\{-\frac{i}{\hbar} \int_0^T \varepsilon_n(t) dt\right\} \cdot \exp\{i\gamma_n(C)\} |\Psi_n\rangle, \quad (11)$$

where the coefficients $\gamma_n(C)$ can be written as

$$\gamma_n(C) = -Im \oint_C ds \cdot \sum_{m \neq n} \frac{\langle n | \nabla_R \hat{H}(\mathbf{R}) | m \rangle \times \langle m | \nabla_R \hat{H}(\mathbf{R}) | n \rangle}{(\varepsilon_m(\mathbf{R}) - \varepsilon_n(\mathbf{R}))^2}, \quad (12)$$

where $\mathbf{R} = (F_x(t), F_y(t))$ and ds is the total area enclosed by the dots in one complete adiabatic rotation in the 2D plane at the heterojunction. Here ε_m and ε_n correspond to the eigenvalues of (7) associated to the quantum states $|m\rangle$ and $|n\rangle$.

COMPUTATIONAL METHOD

We suppose that a QD is formed in the plane of a two dimensional electron gas of $400 \times 400 \text{ nm}^2$ geometry. The in-plane oscillating fields $F_x(t)$ and $F_y(t)$ is varied in such a way that the QD is transported in a closed loop of circular trajectory (see Fig. 3). To find the Berry phase by an explicit numerical method, we diagonalize the total Hamiltonian $H(t)$ at any fixed time using the Finite Element Method. In particular, We utilize the UMFPAK solver in the COMSOL multiphysics package (n.d.) to find the eigenvalues and eigenfunctions of the two coupled eigenvalue partial differential equation (7). The geometry contains 24910 elements. Since the geometry is much larger compared to the actual lateral size of the QD, we impose Dirichlet boundary conditions. Error vs iteration number shows the convergence of simulations is good.

RESULTS AND DISCUSSIONS

In Fig. 1, we have plotted the modeling results of ground and first excited states wavefunctions squared of GaAs quantum dots with no magnetic and no time dependent distortion potential. In Fig. 3, we use the distortion potential ($f(t)$) as a time dependent function and allow the dot to move adiabatically in a closed loop in the 2D plane. Realistic electron wavefunctions of the dots at different locations ($\theta = 0, \pi/2, \pi, 3\pi/2$) in the 2D plane are shown.

Based on FEM (n.d.), we solved the two coupled time dependent Schrödinger equations (7) with the initial condition $H(x, y, 0)\psi(x, y, 0) = \varepsilon\psi(x, y, 0)$ in the fixed time interval $\theta = [0 : 0.1 : 2\pi]$. The adiabatic theorem guarantees that $\psi'_\theta(x, y, \theta) = 0$. We plotted the evolution of the spin dynamics during the adiabatic movement of the QDs in the 2D plane in Fig. 4. Even in the presence of Zeeman energy, where the magnetic field is applied along z-direction, the spin components in the ground state of the QDs are not well defined due to the presence of spin-orbit couplings (Bednarek et al. 2008, Bednarek & Szafran 2008). It means, $\langle \sigma_z \rangle$ is either 1 or -1 depending on the g-factor of electron in QDs and the components of $\sigma_i (i = x, y)$ varies during the adiabatic movement of the QDs in the 2D plane. Fluctuations in $\langle \sigma_z \rangle$ can be made at degenerate sublevels where g-factor exactly vanishes. In this case, rather than finding a scalar Berry phase, one needs to find the matrix Berry phase acting on the initial states within the subspace of degeneracy. (Prabhakar et al. 2010) Since the motivation of the paper is to investigate the influence of electric field on the scalar Berry phase, we choose the parameters in Fig. 4 in such a way that the g-factor is negative and $\langle \sigma_z \rangle = +1$ (Prabhakar & Reynolds 2009). For g-factor control in quantum dots, see Refs. (Prabhakar & Reynolds 2009, Prabhakar et al. 2011). If one choses $\ell_0 = 40 \text{ nm}$, it can be found that the g-factor is positive and $\langle \sigma_z \rangle = -1$. Depending on the choice of the parameters, one can construct the quantum gates (Hadamard, OR, Controlled NOT gates) with the application of the gate controlled electric fields (Bednarek et al. 2012). For example, when all the spin components are equal to unity, one can have Hadamard gates. Since spin components decay with different phase (see Fig. 4) but they all vanishes at certain degree of orientation in the Bloch sphere, one can find the controlled NOT gates. (Bednarek et al. 2008, Bednarek & Szafran 2008) Also, the transport of the QDs are carried out adiabatically, one can find the similar type of

evolution of the spin dynamics (Fig. 4) in each cycle of rotation which is another efficient way to construct the quantum gates from QDs. Since the periodicity of the propagating waves is different for the pure Rashba and for the pure Dresselhaus case, we see the superposition effect in the x- and y-components of the electron spin in QDs (see Fig. 4).

We now turn to the results associated to the Berry phase that is accumulated during the adiabatic transport of the dots in the 2D plane.

In Fig. 5, we plot the characteristics of the Berry phase vs spin-orbit coupling length. As can be seen, the Berry phase for the pure Rashba and pure Dresselhaus cases are well separated at smaller values of the SO lengths due to the presence of the Rashba case and the Dresselhaus spin-orbit coupling case. At large values of spin-orbit lengths $\lambda_R = \lambda > 1.8\mu\text{m}$, the Berry phases for the pure Rashba and for the pure Dresselhaus spin-orbit coupling cases meet each other because extremely weak spin-orbit coupling coefficients are unable to break the in-plane rotational symmetry. Note that the spin-orbit length is characterized by the applied electric field along the z-direction (see inset plot in Fig. 5) (Prabhakar et al. 2013).

CONCLUSIONS

To conclude, based on the developed mathematical model, we have analyzed the wavefunctions of electrons in QDs during the adiabatic movement of the dots along the circular trajectory. By using the Finite Element Method, we have calculated the evolution of the spin dynamics and shown that the superposition effect can be observed during the adiabatic movement of the QDs in the 2D plane. We have shown that the Berry phase for the pure Rashba and pure Dresselhaus cases are well separated at smaller values of the SO lengths due to the presence of large Rashba-Dresselhaus spin-orbit coupling coefficients.

ACKNOWLEDGEMENTS

This work was supported by NSERC and CRC programs, Canada. RM thanks colleagues at Mevlana University for their hospitality during his visit there. RM and AS are also grateful to TUBITAK for its support.

REFERENCES

- (n.d.). Comsol Multiphysics version 3.5a (www.comsol.com).
- Aleiner, I. L. & Fal'ko, V. I. (2001), 'Spin-orbit coupling effects on quantum transport in lateral semiconductor dots', *Phys. Rev. Lett.* **87**, 256801.
- Bason, M. G., Viteau, M., Malossi, N., Huillery, P., Arimondo, E., Ciampini, D., Fazio, R., Giovannetti, V., Mannella, R. & Morsch, O. (2012), 'High-fidelity quantum driving', *Nat. Phys.* **8**, 147.
- Bednarek, S., owski, J. P. & Skubis, A. (2012), 'Manipulation of a single electron spin in a quantum dot without magnetic field', *Applied Physics Letters* **100**(20), 203103.
- Bednarek, S. & Szafran, B. (2008), 'Spin rotations induced by an electron running in closed trajectories in gated semiconductor nanodevices', *Phys. Rev. Lett.* **101**, 216805.
- Bednarek, S., Szafran, B., Dudek, R. J. & Lis, K. (2008), 'Induced quantum dots and wires: Electron storage and delivery', *Phys. Rev. Lett.* **100**, 126805.

- Berger, S., Pechal, M., Pugnetti, S., Abdumalikov, A. A., Steffen, L., Fedorov, A., Wallraff, A. & Filipp, S. (2012), ‘Geometric phases in superconducting qubits beyond the two-level approximation’, *Phys. Rev. B* **85**, 220502.
- Berry, M. V. (1984), ‘Quantal phase factors accompanying adiabatic changes’, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **392**(1802), 45–57.
- Bychkov, Y. A. & Rashba, E. I. (1984), ‘Oscillatory effects and the magnetic susceptibility of carriers in inversion layers’, *J. Phys. C* **17**, 6039.
- Das Sarma, S., Freedman, M. & Nayak, C. (2005), ‘Topologically protected qubits from a possible non-abelian fractional quantum hall state’, *Phys. Rev. Lett.* **94**, 166802.
- Dresselhaus, G. (1955), ‘Spin-orbit coupling effects in zinc blende structures’, *Phys. Rev.* **100**(2), 580–586.
- Eric Yang, S.-R. (2006), ‘Degenerate states of narrow semiconductor rings in the presence of spin-orbit coupling: Role of time-reversal and large gauge transformations’, *Phys. Rev. B* **74**(7), 075315.
- Hu, X. & Das Sarma, S. (2000), ‘Hilbert-space structure of a solid-state quantum computer: Two-electron states of a double-quantum-dot artificial molecule’, *Phys. Rev. A* **61**, 062301.
- Leek, P. J., Fink, J. M., Blais, A., Bianchetti, R., Goppl, M., Gambetta, J. M., Schuster, D. I., Frunzio, L., Schoelkopf, R. J. & Wallraff, A. (2007), ‘Observation of berry’s phase in a solid-state qubit’, *Science* **318**(5858), 1889–1892.
- Loss, D., Goldbart, P. & Balatsky, A. V. (1990), ‘Berry’s phase and persistent charge and spin currents in textured mesoscopic rings’, *Phys. Rev. Lett.* **65**, 1655–1658.
- Nayak, C., Simon, S. H., Stern, A., Freedman, M. & Das Sarma, S. (2008), ‘Non-abelian anyons and topological quantum computation’, *Rev. Mod. Phys.* **80**, 1083–1159.
- Pechal, M., Berger, S., Abdumalikov, A. A., Fink, J. M., Mlynek, J. A., Steffen, L., Wallraff, A. & Filipp, S. (2012), ‘Geometric phase and nonadiabatic effects in an electronic harmonic oscillator’, *Phys. Rev. Lett.* **108**, 170401.
- Prabhakar, S., Melnik, R. & Bonilla, L. L. (2013), ‘Electrical control of phonon-mediated spin relaxation rate in semiconductor quantum dots: Rashba versus dresselhaus spin-orbit coupling’, *Phys. Rev. B* **87**, 235202.
- Prabhakar, S. & Raynolds, J. E. (2009), ‘Gate control of a quantum dot single-electron spin in realistic confining potentials: Anisotropy effects’, *Phys. Rev. B* **79**, 195307.
- Prabhakar, S., Raynolds, J. E. & Melnik, R. (2011), ‘Manipulation of the landé g factor in InAs quantum dots through the application of anisotropic gate potentials: Exact diagonalization, numerical, and perturbation methods’, *Phys. Rev. B* **84**, 155208.
- Prabhakar, S., Raynolds, J., Inomata, A. & Melnik, R. (2010), ‘Manipulation of single electron spin in a GaAs quantum dot through the application of geometric phases: The feynman disentangling technique’, *Phys. Rev. B* **82**, 195306.
- San-Jose, P., Scharfenberger, B., Schön, G., Shnirman, A. & Zarand, G. (2008), ‘Geometric phases in semiconductor spin qubits: Manipulations and decoherence’, *Phys. Rev. B* **77**, 045305.
- Tserkovnyak, Y. & Loss, D. (2011), ‘Universal quantum computation with ordered spin-chain networks’, *Phys. Rev. A* **84**, 032333.
- Wang, H. & Zhu, K.-D. (2008), ‘Voltage-controlled berry phases in two vertically coupled InGaAs/GaAs quantum dots’, *EPL* **82**(6), 60006.
- Wilczek, F. & Zee, A. (1984), ‘Appearance of gauge structure in simple dynamical systems’, *Phys. Rev. Lett.* **52**(24), 2111–2114.
- Wu, Y., Piper, I. M., Ediger, M., Brereton, P., Schmidgall, E. R., Eastham, P. R., Hugues, M., Hopkinson, M. & Phillips, R. T. (2011), ‘Population inversion in a single InGaAs quantum dot using the method of adiabatic rapid passage’, *Phys. Rev. Lett.* **106**, 067401.
- Xiao, D., Chang, M.-C. & Niu, Q. (2010), ‘Berry phase effects on electronic properties’, *Rev. Mod. Phys.* **82**, 1959–2007.
- Yang, S.-R. E. (2007), ‘Control of many-electron states in semiconductor quantum dots by non-abelian vector potentials’, *Phys. Rev. B* **75**(24), 245328.
- Yang, S.-R. E. & Hwang, N. Y. (2006), ‘Single electron control in n-type semiconductor quantum dots using non-abelian holonomies generated by spin orbit coupling’, *Phys. Rev. B* **73**(12), 125330.

EFFICIENT MODELING OF GRAPHENE BASED OPTICAL DEVICES

Costantino De Angelis,
Andrea Locatelli
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Brescia
Via Branze 38, 25123 Brescia, Italy
Email: costantino.deangelis@unibs.it

KEYWORDS

Electromagnetic modeling, two-dimensional materials, graphene, finite-difference techniques, Beam Propagation Method, directional couplers.

ABSTRACT

In this paper we describe some basic issues in the numerical modeling of graphene devices at optical frequencies. In the first part of the paper we analytically describe the behavior of a graphene-based directional coupler; in the second part of the paper we introduce a novel formulation of the Beam Propagation Method to show how this powerful numerical technique can also be used to describe light propagation in graphene as well as in other new emerging two-dimensional materials.

COUPLING BETWEEN GRAPHENE POLARITONS

We discuss here the tuning of the coupling of surface plasmon polaritons between two graphene layers with nanometer spacing. We demonstrate that by slightly changing the electrical doping and then shifting the chemical potential, a graphene coupler can switch from the bar to the cross state. As a consequence, the coupling coefficient in such structures can be easily controlled in a un ultrafast fashion by means of an applied electrical signal (Auditore et al. 2013b). These findings open the way to fully exploit the huge nonlinearity of graphene for all optical signal processing: from one side giving more degrees of freedom to already proposed devices (Auditore et al. 2013a; Smirnova et al. 2013; Buslaev et al. 2013; Locatelli et al. 2014; Bludov et al. 2014), from the other side paving the way to new devices. Here in particular we want to use this very interesting structure as a benchmark where to exploit the numerical technique we will describe in the next section.

As well known, graphene can sustain surface plasmon polaritons (SPP) having unique properties as compared to what we are used to with noble metals. In fact a single layer of graphene can support either TE or TM polarized plasmons without suffering from huge loss (Mikhailov et al. 2007; Mikhailov et al. 2008; Jablan et al. 2009); moreover, as far as TM polarization is concerned, the extremely high

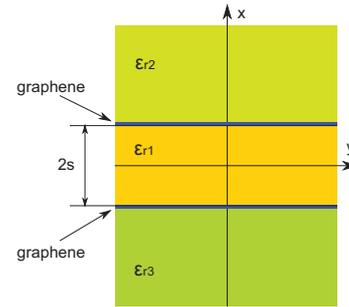


Fig. 1. Schematic of the graphene directional coupler: the separation between the layers is equal to $2s$.

confinement factor is particularly favourable to explore the huge $\chi^{(3)}$ nonlinearity of graphene (Auditore et al. 2013a; Zhang et al. 2012; Gorbach 2013). Experimental endeavors have demonstrated the evidences of the existence of graphene plasmons by measuring the plasmon resonance of graphene nanoribbon arrays (Geng et al. 2011), and by acquiring their near field images (Chen et al. 2012). The coupling of SPP between separated graphene layers has been recently analyzed in (Wang et al. 2012); however the very interesting properties arising from the easily tunable optical properties of graphene have not been exploited yet in this framework. Here we show that by slightly changing the chemical potential, a graphene coupler can switch from the bar to the cross state; as a consequence, the output state in such structures can be easily controlled by means of an applied electrical signal.

In Figure (1) we report the basic geometry we consider in this section; two graphene layers are embedded in a dielectric structure: region 1 (of width $2s$) is the dielectric in between the two graphene layers. At the graphene boundary we set the following conditions on the tangential components of the electromagnetic field:

$$\begin{aligned} (\vec{E}_{2,3} - \vec{E}_1) \times \hat{x} &= 0 \\ (\vec{H}_{2,3} - \vec{H}_1) \times \hat{x} &= \pm i\omega\epsilon_0\epsilon_{rS1-2,3}\vec{E}_{||}(x = \pm s), \end{aligned} \quad (1)$$

where $\vec{E}_{||}$ is the electric field tangent to the graphene layer and ϵ_{rS1-2} (ϵ_{rS1-3}) is the relative surface permittivity of the

graphene layer between regions 1 and 2 (3). As far as the electromagnetic constants of graphene are concerned, we write the linear contribution to the relative complex permittivity as (Vakil et al. 2011; Stauber et al. 2008):

$$\epsilon_{rC} = \frac{\epsilon_{rS}}{d_g} = 1 + \frac{\sigma_{\Sigma,I}^{(1)}}{d_g \omega \epsilon_0} - i \frac{\sigma_{\Sigma,R}^{(1)}}{d_g \omega \epsilon_0} = \epsilon_{rC,R} + i \epsilon_{rC,I} \quad (2)$$

where d_g is the graphene thickness and the surface complex conductivity $\sigma_{\Sigma}^{(1)} = \sigma_{\Sigma,R}^{(1)} + i \sigma_{\Sigma,I}^{(1)}$ (in Siemens) is obtained from theoretical models now well established and experimentally validated (Stauber et al. 2008), which give the following dependence of the real and imaginary parts of the conductivity on frequency (ω), temperature (T) and chemical potential (μ_c):

$$\begin{aligned} \sigma_{\Sigma,R}^{(1)}(\omega) &\simeq \frac{\sigma_0}{2} \left(\tanh \frac{\hbar\omega + 2\mu_c}{4k_B T} + \tanh \frac{\hbar\omega - 2\mu_c}{4k_B T} \right) \\ \sigma_{\Sigma,I}^{(1)}(\omega) &\simeq \frac{\sigma_0}{\pi} \left[\frac{4}{\hbar\omega} \left(\mu_c - \frac{2\mu_c^3}{9t^2} \right) - \log \frac{\hbar\omega + 2\mu_c}{\hbar\omega - 2\mu_c} \right] \end{aligned} \quad (3)$$

where $t = 2.7$ eV is the hopping parameter, \hbar and k_B are the reduced Planck's and Boltzmann's constants, respectively, and $\sigma_0 = e^2/(4\hbar) \simeq 6.08 \cdot 10^{-5}$ S, with e the electron charge. Note that when graphene is modeled as a brick with a finite thickness ($d_g = 0.34$ nm), the results reported in Equation (3) (where we have conductivities with the dimensions of Siemens) need to be normalized as follows: $\sigma \rightarrow \sigma/d_g$ to get conductivities in Siemens/m.

Note also that this model can be easily extended into the nonlinear regime by adding a nonlinear correction to the surface conductivity as: $\sigma_{\Sigma} = \sigma_{\Sigma}^{(1)} + \sigma_{\Sigma}^{(3)} |\vec{E}_{\parallel}|^2$ (Mikhailov et al. 2008). Moreover, thanks to the extremely small thickness of the graphene layer, nonlinearity can be analyzed by a parameter embedded into the coefficients describing the continuity of the tangential components of the electromagnetic field (Auditore et al. 2013a; Gorbach 2013).

To describe SPP propagation along z , we first note that, at first order, the y dependence of the electromagnetic field can be neglected; we then look for guided modes with harmonic temporal dependence $\exp(i\omega t)$ and spatial variation $\vec{E}_{1,2,3}(x, z)$, $\vec{H}_{1,2,3}(x, z) \sim \exp(-i\beta z \pm \Gamma_{1,2,3} x)$ with $\Gamma_{1,2,3}^2 = \beta^2 - \epsilon_{r1,2,3} k_0^2$. Obviously the complex wavenumber β , through its real and imaginary parts, describes the evolution of the phase and the amplitude of the guided modes. We can apply the above modeling to derive the dispersion relation of both TE and TM modes. In the following we describe in detail the TM polarization. We first consider a very general situation where the two graphene layers can be biased in a different way to give rise to an asymmetric coupler. We assume that the mode profiles exhibit a behavior which is exponentially decaying from the graphene layers, as shown in (Wang et al. 2012), and we evaluate the proper boundary conditions at $x = \pm s$ (see Equations 1). After straightforward algebra we find that coupled SPP in the system are determined by setting to zero the determinant of the following matrix of coefficients:

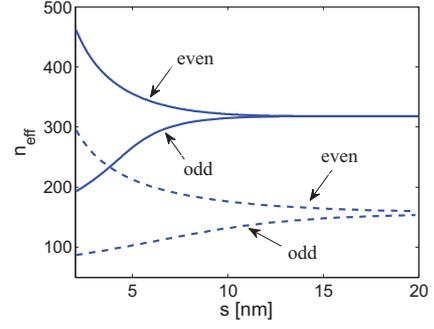


Fig. 2. Effective index $n_{eff} = \Re(\beta)/k_0$ of even and odd supermodes of the coupled graphene layers as a function of the separation among the layers. Continuous (dashed) lines refer to a chemical potential of $\mu_c = 0.1$ eV ($\mu_c = 0.15$ eV).

$$M = \begin{bmatrix} e^{\Gamma_1 s} & e^{-\Gamma_1 s} & -e^{-\Gamma_2 s} & 0 \\ e^{-\Gamma_1 s} & e^{\Gamma_1 s} & 0 & -e^{-\Gamma_3 s} \\ \frac{i\omega\epsilon_1}{\Gamma_1} e^{\Gamma_1 s} & -\frac{i\omega\epsilon_1}{\Gamma_1} e^{-\Gamma_1 s} & g_{1-2} e^{-\Gamma_2 s} & 0 \\ -\frac{i\omega\epsilon_1}{\Gamma_1} e^{-\Gamma_1 s} & \frac{i\omega\epsilon_1}{\Gamma_1} e^{\Gamma_1 s} & 0 & g_{1-3} e^{-\Gamma_3 s} \end{bmatrix} \quad (4)$$

where g_{1-2} and g_{1-3} take into account the contribution of the two graphene layers in the continuity conditions:

$$g_{1-2} = i\omega\epsilon_0\epsilon_{rS,1-2} + \frac{i\omega\epsilon_2}{\Gamma_2}, \quad g_{1-3} = i\omega\epsilon_0\epsilon_{rS,1-3} + \frac{i\omega\epsilon_3}{\Gamma_3}$$

where $\epsilon_{rS,1-2}$ and $\epsilon_{rS,1-3}$ refer to the relative surface dielectric constant of the two graphene layers, which can in general get different values due to different carriers concentrations. This asymmetric coupler offers a wide variety of possible settings which certainly deserve to be investigated both in the linear and in the non linear regime. Here we want to give a first prototype example into the possibilities offered by the tunability of the graphene parameters in this framework; we thus focus our attention on a very particular situation corresponding to a linear and symmetric case ($\epsilon_2 = \epsilon_3$ and $\epsilon_{rS,1-2} = \epsilon_{rS,1-3}$); moreover we use $T = 300$ K, $\lambda = 10$ μm and for the sake of simplicity we also set $\epsilon_1 = \epsilon_2 = \epsilon_3 = 2.25$. In this regime the graphene directional coupler has two different eigenstates: the even (odd) supermode corresponding to the out-of-phase (in-phase) hybridization of the SPP guided by the single graphene layer. Notice also that the even mode here always has the higher value of the propagation constant. In Figure (2) we report the solution of the dispersion relations as a function of s for two different situations: continuous lines here refer to the even and odd supermodes corresponding to a chemical potential $\mu_c = 0.1$ eV in Equations (3), while the dashed lines refer to a choice of the chemical potential $\mu_c = 0.15$ eV in Equations (3). It is straightforward to understand that for large enough s the two supermodes of the coupler tend to degeneracy and their propagation constant goes into the propagation constant of the SPP of the single graphene layer. The main message we can read from Figure (2) is that

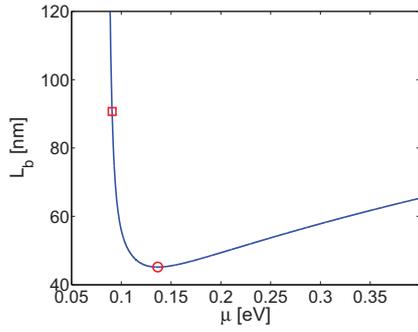


Fig. 3. Beat length versus chemical potential for a graphene plasmon coupler. Here $2s = 10$ nm.

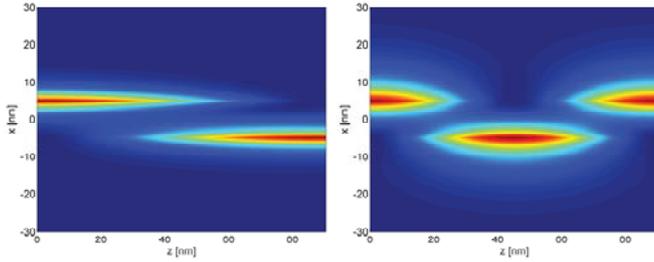


Fig. 4. Field evolution in a graphene plasmon directional coupler: left (right) refers to a chemical potential of $\mu_c = 0.09$ eV ($\mu_c = 0.14$ eV). Here $2s = 10$ nm.

a very small change in the chemical potential (i.e. a very small change of an electrical signal applied to the graphene layers) can induce a very big change in the behaviour of the system. In the following we focus our attention at $s = 5$ nm. For this value of the separation between the layers we have computed the beat length of the directional coupler as a function of the chemical potential and the results are reported in Figure (3). We can clearly see there that a very small change of the chemical potential can be used to induce huge changes of the beat length of the coupler. The two particular points (open square and open circle) enlightened in Figure (3) are the initial conditions in Figure (4), where we describe the propagation of the electromagnetic signal in the graphene coupler. In both panels in Figure (4) total propagation length is set to $L \simeq 90$ nm. On the left panel in Figure (4) the input condition corresponds to the square in Figure (3) and the coupler is in the cross state; on the right panel the input condition corresponds to the circle in Figure (3) and the coupler is in the bar state.

NOVEL BPM FOR GRAPHENE-BASED DEVICES

In this section we describe a finite-difference Beam Propagation Method (BPM) capable of dealing with the discontinuity of the tangential component of the magnetic field induced by two-dimensional graphene layers which can be arbitrarily placed within dielectric media. In stark contrast with conventional numerical solvers, this approach does not require a discretization step as small as a fraction of the atomic thickness of graphene, allowing ultra-fast simulation times. The validity of the method is proved by propagating

the plasmonic supermodes of two coupled graphene layers, and the evaluated beat length exhibits excellent agreement with respect to analytical results obtained by using the procedure we have illustrated in the previous section.

In the last decades the Beam Propagation Method, originally conceived by Feit and Fleck in (Feit et al. 1978), has been one of the most employed numerical techniques for the analysis of electromagnetic wave propagation in optical devices. Since its first formulation many improvements have been made, allowing for fast and accurate semi- and full-vectorial analysis of high index contrast and graded index waveguides (Huang et al. 1993), taking into account also backward propagating waves in multilayered structures both in linear and nonlinear regimes (Locatelli et al. 2002).

The recent advances in theoretical and experimental studies concerning graphene have opened up new scenarios in the field of photonics, envisaging the birth of novel applications which exploit the peculiar properties of this fascinating material (Bao et al. 2012). In this context, numerical modeling plays a crucial role since accurate results are required in short computational times. As a matter of fact, existing full-wave numerical solvers can still be used, as long as graphene is modeled as a sub-nanometer layer (Locatelli et al. 2012). On the one hand this solves the problem of emulating the two-dimensional nature of graphene, on the other hand discretization steps smaller than its atomic thickness are required, leading to an unacceptable computational burden.

In the first part of the paper we have demonstrated that the discontinuity of the tangential magnetic field component induced by the currents flowing on a two-dimensional graphene layer can be treated as a boundary condition instead of considering a volumetric layer of graphene in the problem. This technique allowed to derive dispersive properties and mode profiles of graphene plasmon waveguides (Auditore et al. 2013b), and it was used in combination with an asymptotic expansion of Maxwell equations to derive amplitude equations for surface plasmon waves (Smirnova et al. 2013; Gorbach 2013).

Here we describe a finite-difference formulation of the second derivative of the tangential component of the magnetic field, which exhibits discontinuities due to the surface currents flowing on the graphene layers. We exploit a methodology, first applied by Stern for the discretization of the discontinuous electric field of a TM mode (Stern 1998), which allows us to include all the effects induced by the graphene in a reformulated diffractive operator, without resorting to sub-nanometer sampling steps. The algorithm has been validated by propagating the supermodes of two coupled graphene layers, and by comparing these data with known results.

We consider planar and piecewise constant dielectric structures where x and z are the transverse and the propagation directions respectively, with no variation in the y coordinate. Under these assumptions and focusing our attention on harmonic TM waves, the propagation of the H_y component of the magnetic field obeys the Helmholtz equation. If we express the magnetic field as $H_y(x, z) = \mathcal{H}_y(x, z) \exp(-j\kappa_0 n_0 z)$, where

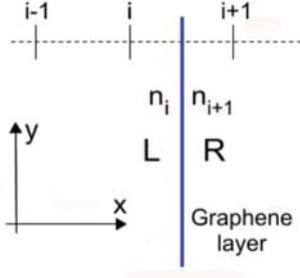


Fig. 5. Schematic view of the transverse discretized domain.

κ_0 is the wave-number in vacuum, and n_0 is a reference refractive index (Feit et al. 1978), Helmholtz equation transforms to:

$$\frac{\partial^2 \mathcal{H}_y}{\partial z^2} - 2j\kappa_0 n_0 \frac{\partial \mathcal{H}_y}{\partial z} + \frac{\partial^2 \mathcal{H}_y}{\partial x^2} + \kappa_0^2 (n^2(x) - n_0^2) \mathcal{H}_y = 0, \quad (5)$$

where $n(x)$ is the refractive index profile. By neglecting the first term on the left-hand side of Equation (5), the well-known Schrödinger equation is obtained.

In the absence of any graphene layer, \mathcal{H}_y is continuous across the transverse domain, while a discontinuity of $\partial \mathcal{H}_y / \partial x$ occurs at each dielectric interface. If a sheet of graphene is introduced within the geometry (see Figure (5)), a discontinuity of the tangential component of the magnetic field is induced by the surface currents flowing on the graphene layer:

$$\left(\vec{H} \Big|_R - \vec{H} \Big|_L \right) \times \hat{x} = j\omega \epsilon_0 \epsilon_{gr\Sigma} \vec{E}_{\parallel} \Big|_{R,L}, \quad (6)$$

where the relative surface dielectric constant of graphene is:

$$\epsilon_{gr\Sigma} = 1 + \frac{\sigma_{\Sigma,I}}{\omega \epsilon_0} - j \frac{\sigma_{\Sigma,R}}{\omega \epsilon_0}, \quad (7)$$

with $\sigma_{\Sigma} = \sigma_{\Sigma,R} + j\sigma_{\Sigma,I}$ being the surface conductivity of graphene (Stauber et al. 2008), subscripts R and L refer to right and left graphene boundaries, and \vec{E}_{\parallel} is the electric field in the plane of the graphene layer. In our case Equation (6) reads as:

$$\mathcal{H}_y \Big|_R - \mathcal{H}_y \Big|_L = j\omega \epsilon_0 \epsilon_{gr\Sigma} \mathcal{E}_z \Big|_{R,L}. \quad (8)$$

Moreover, whenever a dielectric discontinuity occurs, the following equation derived from Maxwell equations holds at the corresponding boundary:

$$\frac{\partial \mathcal{H}_y}{\partial x} \Big|_R = \frac{n^2 \Big|_R}{n^2 \Big|_L} \frac{\partial \mathcal{H}_y}{\partial x} \Big|_L. \quad (9)$$

In order to model the effects of graphene-induced surface currents and dielectric discontinuities, special care must be devoted to the finite-difference implementation of the second derivative with respect to the transverse coordinate x , since samples fall before and after a graphene layer or a dielectric discontinuity. Following the procedure described by Stern in the Appendix of (Stern 1998), and labeling for the sake of simplicity \mathcal{H}_y as H , and \mathcal{E}_z as E , the finite-difference approximation of the second derivative of the magnetic field

component, evaluated at point i with a dielectric interface or a graphene layer in-between points i and $i+1$ (see Figure (5)) can be written as:

$$\frac{\partial^2 H}{\partial x^2} \Big|_i = \frac{H_{i-1} - 2H_i + H_{i+1}^*}{\Delta x^2}, \quad (10)$$

where Δx is the transverse step and the superscript $*$ is introduced to indicate that a discontinuity between points i and $i+1$ must be taken into account. Now using the Taylor series expansion for H and the finite-difference approximation for its first derivative we can get:

$$H \Big|_R = \frac{H_{i+1} + H_i^*}{2}, \quad H \Big|_L = \frac{H_{i+1}^* + H_i}{2} \quad (11)$$

$$\frac{\partial H}{\partial x} \Big|_R = \frac{H_{i+1} - H_i^*}{\Delta x}, \quad \frac{\partial H}{\partial x} \Big|_L = \frac{H_{i+1}^* - H_i}{\Delta x}. \quad (12)$$

By substituting Equation (11) in Equation (8), and Equation (12) in Equation (9) we can write that:

$$(H_{i+1} + H_i^*) = (H_{i+1}^* + H_i) + 2\Delta E \Big|_L \quad (13)$$

$$(H_{i+1} - H_i^*) = \theta_i (H_{i+1}^* - H_i), \quad (14)$$

where we defined $\Delta = j\omega \epsilon_0 \epsilon_{gr\Sigma}$ and $\theta_i = n_{i+1}^2 / n_i^2$. From Equation (13) $H_i^* = H_{i+1} + H_i - H_{i+1} + 2\Delta E \Big|_L$, which can be substituted in Equation (14) to give:

$$H_{i+1}^* = \frac{2}{(\theta_i + 1)} H_{i+1} + \frac{(\theta_i - 1)}{(\theta_i + 1)} H_i - \frac{2\Delta}{(\theta_i + 1)} E \Big|_L. \quad (15)$$

From Maxwell equations we can write:

$$\mathcal{E}_z \Big|_L = -j \frac{1}{\omega \epsilon_0 n^2 \Big|_L} \frac{\partial \mathcal{H}_y}{\partial x} \Big|_L, \quad E \Big|_L = -j \frac{1}{\omega \epsilon_0 n_i^2} \frac{(H_{i+1}^* - H_i)}{\Delta x}, \quad (16)$$

which can be inserted in Equation (15) and allows to formulate Equation (10) in terms of “regular” H_{i-1} , H_i , and H_{i+1} samples. Analogously, for a dielectric discontinuity or a graphene layer in-between points i and $i-1$, the finite-difference approximation of the second derivative at point i can be written as:

$$\frac{\partial^2 H}{\partial x^2} \Big|_i = \frac{H_{i-1}^{**} - 2H_i + H_{i+1}}{\Delta x^2}, \quad (17)$$

where the superscript $**$ indicates that a discontinuity between points i and $i-1$ must be taken into account. By following the same procedure we used above, it is possible to formulate Equation (17) in terms of “regular” H_{i-1} , H_i , and H_{i+1} samples.

In the case of discontinuities in-between points i and $i+1$ and in-between points i and $i-1$, Equations (10) and (17) merge giving the final finite-difference formulation of the second derivative of the magnetic field evaluated at position i (Stern 1998):

$$\frac{\partial^2 H}{\partial x^2} \Big|_i = \frac{H_{i-1}^{**} - 2H_i + H_{i+1}^*}{\Delta x^2}, \quad (18)$$

which can be written in terms of the “regular” samples weighted with coefficients depending on the refractive indexes and on the relative surface dielectric constant of graphene.

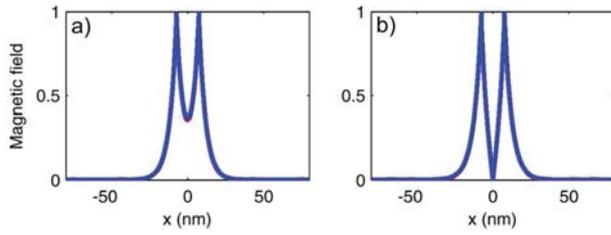


Fig. 6. Magnitude of the magnetic field at the input of the coupler (blue lines) and after propagation with the BPM (red lines). a) Even supermode. b) Odd supermode. The graphene layers are placed near the peaks of the curves.

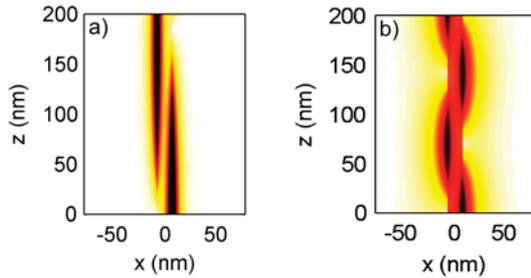


Fig. 7. Magnitude of the magnetic field along the graphene coupler. a) $\mu_c = 0.1$ eV. b) $\mu_c = 0.16$ eV. Beating between the two layers is clearly visible.

To verify the soundness of our method we have analyzed the graphene directional coupler which was described in the previous section, since a benchmark structure with known properties is very helpful in order to validate the novel numerical technique. Two graphene monolayers are placed with a spacing d among them. These are immersed in a uniform dielectric medium with relative dielectric constant $\epsilon_s = 2.25$, and a bias voltage is applied between the sheets with the goal of tuning the two-dimensional complex conductivity σ_Σ of graphene, which is fixed to the same value for both the layers because of the peculiar characteristics of the Dirac dispersion relation of electrons in this material (Stauber et al. 2008; Bao et al. 2012). It is worth noting that the dependence of σ_Σ on the chemical potential μ_c (and then on the bias voltage) has been evaluated from well-established analytical models (Stauber et al. 2008). In order to verify the accuracy of our algorithm we have compared BPM results with the solution of the dispersion relations of the two supermodes obtained by using the procedure reported in the previous section, which allows propagation constants and mode profiles to be calculated.

First, we analytically evaluated the supermode profiles when $d = 15$ nm and $\mu_c = 0.1$ eV, and we separately launched them into the coupler for a propagation length equal to 200 nm. In Figure (6) we depict the magnitude of the magnetic field of the single supermodes at the input of the structure and after BPM propagation. Input and output curves are almost perfectly overlapped, and this is a first proof of the effectiveness of our algorithm. Moreover, other results not reported here show that when the input field is slightly perturbed propagation correctly transforms it into the analytical mode of the coupler.

Then, we excited the graphene coupler with the superpo-

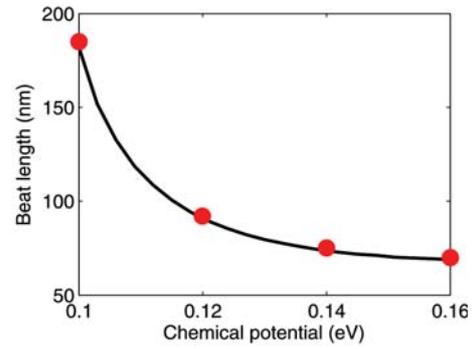


Fig. 8. Beat length as a function of the chemical potential evaluated from the BPM (red circles) and from the analytical formulas (black solid line).

sition of the supermodes in order to evaluate its beat length. We report in the panel (a) of Figure (7) the magnitude of the magnetic field along the coupler with $\mu_c = 0.1$ eV, and we compare the beat length which can be estimated from these data with the value coming from modal analysis through the well-known relation $L_b = \pi / (\beta_{even} - \beta_{odd})$. Different methods exhibit very good agreement, indeed the beat length is around 185 nm according to the BPM simulation, whereas the modal analysis predicts that L_b should be equal to 182 nm. We changed the chemical potential to $\mu_c = 0.16$ eV and we repeated the same analysis, by obtaining the results illustrated in the panel (b) of Figure (7). Notice that a slight change of the bias voltage induces a huge modification of the behavior of the coupler, in fact the BPM algorithm estimates that in this case L_b should decrease to 70 nm, which is in excellent agreement with the result coming from the dispersion relations of the supermodes ($L_b = 69$ nm).

The dependence of the beat length L_b on the chemical potential μ_c has been systematically investigated by using our BPM technique, always comparing these results with the reference solution provided by the dispersion relations of the supermodes. In Figure (8) the beat length as a function of the chemical potential is depicted when $d = 15$ nm and μ_c is varied between 0.1 and 0.16 eV. Four BPM simulations characterized by different values of μ_c (and then by different conductivity of the graphene layers) have been performed, with the input field which was the sum of the two supermodes at the corresponding chemical potential. It is straightforward to see that light propagation performed by using our BPM technique predicts a field evolution inside the benchmark structure (Auditore et al. 2013b; Smirnova et al. 2013) which is in excellent agreement with the analytical estimates of L_b , and this constitutes a strong validation of the proposed method.

CONCLUSIONS

In conclusion, in the first part of this paper we have described the tuning of the coupling of surface plasmon polaritons between two spatially separated graphene layers. We were able to prove that, by slightly changing the chemical potential, a graphene coupler can switch from the bar to the cross state; as a consequence, the coupling coefficient

in such structures can be easily controlled by means of an applied electrical signal thus opening interesting opportunities for signal processing. In the second part of the work we have introduced a novel numerical technique which exploits the two-dimensional nature of graphene to greatly reduce the computational burden in the modeling of graphene-based devices. The effectiveness of this technique has been demonstrated by using as a benchmark the analysis of the tunability of the coupling of surface plasmon polaritons between two spatially separated graphene layers.

REFERENCES

- Auditore, A.; C. De Angelis; A. Locatelli; S. Boscolo; M. Midrio; M. Romagnoli; A.-D. Capobianco; and G. Nalesso. 2013. "Graphene sustained nonlinear modes in dielectric waveguides," *Optics Letters* 38, 631–633.
- Auditore, A.; C. De Angelis; A. Locatelli; and A. B. Aceves. 2013. "Tuning of surface plasmon polaritons beat length in graphene directional couplers," *Optics Letters* 38, 4228–4231.
- Bao, Q.; and K. P. Loh. 2012. "Graphene Photonics, Plasmonics, and Broadband Optoelectronic Devices," *ACS Nano* 6, 3677–3694.
- Bludov, Y. V., D. A. Smirnova; Y. S. Kivshar; N. M. R. Peres; and M. I. Vasilevskiy. 2014. "Nonlinear TE-polarized surface polaritons on graphene," *Physical Review B* 89, 035406(1–6).
- Buslaev, P. I.; I. V. Iorsh; I. V. Shadrivov; P. A. Belov; and Y. S. Kivshar. 2013. "Plasmons in waveguide structures formed by two graphene layers," *JETP Letters* 97, 535–539.
- Chen, J.; M. Badioli; P. Alonso-Gonzalez; S. Thongrattanasiri; F. Huth; J. Osmond; M. Spasenovic; A. Centeno; A. Pesquera; P. Godignon; A. Zurutuza; N. Camara; F. J. Garcia de Abajo; R. Hillenbrand; and F. Koppens. 2012. "Optical nano-imaging of gate-tunable graphene plasmons," *Nature* 487, 77–81.
- Feit, M. D.; and J. A. Fleck. 1978. "Light propagation in graded-index optical fibers," *Applied Optics* 17, 3990–3998.
- Geng, B.; L. Ju; J. Horng; C. Girit; M. Martin; Z. Hao; H. A. Bechtel; X. Liang; A. Zettl; Y. R. Shen; and F. Wang. 2011. "Graphene plasmonics for tunable terahertz metamaterials," *Nature Nanotechnology* 6, 630–634.
- Gorbach, A. V. 2013. "Nonlinear graphene plasmonics: Amplitude equation for surface plasmons," *Physical Review A* 87, 013830(1–7).
- Hadley, G. R. 1992. "Wide-angle beam propagation using Padé approximant operators," *Optics Letters* 17, 1426–1428.
- Huang, W. P.; and C. L. Xu. 1993. "Simulation of Three-Dimensional Optical Waveguides by Full-Vector Beam Propagation Method," *IEEE Journal of Quantum Electronics* 29, 2639–2649.
- Jablan, M.; H. Buljan; and M. Soljacic. 2009. "Plasmonics in graphene at infrared frequencies," *Physical Review B* 80, 245435(1–7).
- Locatelli, A.; F. M. Pigozzo; D. Modotto; A.-D. Capobianco; and C. De Angelis. 2002. "Bidirectional Beam Propagation Method for Multilayered Dielectrics With Quadratic Nonlinearity," *IEEE Journal of Selected Topics in Quantum Electronics* 8, 440–447.
- Locatelli, A.; A.-D. Capobianco; M. Midrio; S. Boscolo; and C. De Angelis. 2012. "Graphene-assisted control of coupling between optical waveguides," *Optics Express* 20, 28479–28484.
- Locatelli, A.; A.-D. Capobianco; G. Nalesso; S. Boscolo; M. Midrio; and C. De Angelis. 2014. "Graphene based electro-optical control of the beat length of dielectric couplers," *Optics Communications* 318, 175–179.
- Mikhailov, S. A.; and K. Ziegler. 2007. "New electromagnetic mode in graphene," *Physical Review Letters* 99, 016803(1–4).
- Mikhailov, S. A.; and K. Ziegler. 2008. "Nonlinear electromagnetic response of graphene: frequency multiplication and the self-consistent-field effects," *Journal of Physics: Condensed Matter* 20, 1–10.
- Smirnova, D. A.; A. V. Gorbach; I. V. Iorsh; I. V. Shadrivov; and Y. S. Kivshar. 2013. "Nonlinear switching with a graphene coupler," *Physical Review B* 88, 045443(1–5).
- Stauber, T.; N. M. R. Peres; and A. K. Geim. 2008. "Optical conductivity of graphene in the visible region of the spectrum," *Physical Review B* 78, 085432(1–8).
- Stern, M. S. 1998. "Semivectorial polarised finite difference method for optical waveguides with arbitrary index profiles," *IEE Proceedings* 135, 56–63.
- Vakil, A.; and N. Engheta. 2011. "Transformation optics using graphene," *Science* 332, 1291–1294.
- Wang, B.; X. Zhang; X. Yuan; and J. Teng. 2012. "Optical coupling of surface plasmons between graphene sheets," *Applied Physics Letters* 100, 131111(1–4).
- Zhang, H.; S. Virally; Q. Bao; L. K. Ping; S. Massar; N. Godbout; and P. Kockaert. 2012. "Z-scan measurement of the nonlinear refractive index of graphene," *Optics Letters* 37, 1856–1858.

Simulation in Industry, Business and Services

A Simulation Study: The Business Value of E-business for a Maintenance Provider

Orit Raphaeli, Liron Rosenfeld, Lior Fink, and Sigal Berman
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
84105, Beer-Sheva, Israel
E-mail: sigalbe@bgu.ac.il

KEYWORDS

Business value of IT, E-business, Supply Chain Management, Discrete Event Simulation.

ABSTRACT

The exploitation of IT using the Internet platform to integrate inter-firm processes, often termed “e-business”, has gained research attention due to its potential to enhance the competitive position of businesses. However, the business value of new e-business technologies has remained unexamined given the complexity of the processes through which value is created. In this study we investigate the value of an emerging e-business technology, founded on mobile cloud computing technology. We focus on evaluating the prospective impacts of a collaborative manufacturing network, currently developed by an EC research project, in an automation maintenance company. Three simulation models were constructed, one for the traditional corrective maintenance operation mode as well as two alternative maintenance scenarios enabled by the collaborative platform that allows dynamic allocation of technicians, based on fault status. The results show that by using the collaborative platform, efficiency and service quality can be improved along with reduced cost and improved sustainability. Moreover, communication infrastructure quality negatively affected system contribution. Results also show that the collaborative capabilities, allowing effective handling of different demand volumes, have also improved system flexibility along the demand dimension. Managerial implications of the results are discussed.

INTRODUCTION

The new supply chain process capabilities enabled by Internet-based Information Technology (IT) have stimulated a shift toward digitized integration across supply chain processes that is gradually replacing the conventional processes between supply chain entities (Dong et al., 2009). This shift emphasizes the exploitation of IT using the Internet platform to integrate inter-firm processes, from upstream (supplier) to downstream (customer) operations (Lee, 2000). Often termed “e-business”, Internet-based supply chain integration (Zhu, 2004) enables the sharing of accurate and timely information and the coordination of activities

between business entities. Such e-business-based capabilities are expected to enhance the competitive position of businesses that successfully incorporate them (Rai et al., 2006). Much research, in both the information systems and operations research disciplines, has focused on the relationship between e-business technologies and organizational performance (Zhu and Kraemer, 2005). However, the value of e-business technologies remained an elusive concept due to contradictory results, attributed mainly to the inconsistent definitions of the main concepts and lack of consideration of the contingency effects of business conditions (Van der Vaart and van Donk, 2008; Zhang et al., 2011). In addition, current literature has primarily focused on retrospective and subjective perspectives, using mainly survey-based methods, leaving the value of new e-business technologies unexamined. One such type is Mobile Cloud Computing (MCC), which refers to the combination of cloud computing and mobile networks, enabling execution of rich mobile applications on a plethora of mobile devices. It has been defined as “a rich mobile computing technology that leverages unified elastic resources of varied clouds and network technologies toward unrestricted functionality, storage, and mobility to serve a multitude of mobile devices anywhere, anytime through the channel of Ethernet or Internet regardless of heterogeneous environments and platforms” (Sanaei et al., 2013). Similar to other emerging technologies, determining the business value of MCC is a challenging endeavor, given the complexity of the processes through which value is created (Fink, 2010).

In this study, we evaluate the prospective impacts of MCC technology in the context of supply chain process capabilities. Specifically, we refer to a collaborative manufacturing network, currently developed by an EC research project (ComVantage project), and designed to provide an Internet-based collaboration space, with a secure access control, shared by all relevant supply chain stakeholders. The ComVantage network is based on mobile apps that shall support users in fast decision making and problem solving, using information from different sources across the organisations that is provided and maintained via Linked Data (www.comvantage.eu). The network is intended to provide a secure inter-organisational collaboration space in various manufacturing environments that are

represented in the research project through several use cases. The business value of the collaborative network for the manufacturers in the fashion industry was studied through the evaluation of the prospective network's implementation effects at an Internet-based fashion retailer (Raphaeli et al., 2013). The results indicated mixed performance impacts of upstream and downstream collaboration. While upstream collaboration facilitated improved efficiency but increased costs, downstream collaboration showed negative effects on both aspects.

This study refers to the maintenance industry, through a case study of an automation maintenance company specializing in maintenance of industrial machines. Maintenance of industrial machines is a complex and cost intensive task. It is concerned with providing immediate and efficient service by highly skilled and well trained service personnel. In order to increase its competitive advantage, the automation maintenance company aims at improving preparation of on-site maintenance operations through better identification and assessment of machine faults by introducing the ComVantage collaborative platform.

In the current study, we focus on examining the performance impacts of alternative corrective maintenance processes enabled by ComVantage capabilities to provide real time access to customers' machine data (downstream collaboration). Specifically, we investigate the performance impacts of both remote diagnosis and remote repair capabilities. Since machine data is sensitive, safety plays a major role in the implementation, thus access permission errors are expected. We thus additionally examine the influence of problems in attaining access permission on performance. We also consider the impact of demand characteristics on the process of value creation. The study aims to answer two key questions: (a) How does downstream collaboration affect operational performance? (b) Do demand characteristics influence this relationship?

CASE STUDY DESCRIPTION

iAutomation is an automation maintenance company, specializing in maintenance of industrial machines. iAutomation has 15 customers with either grinding (GM) or spinning (SM) machines. iAutomation employs three types of employees: Mobile Maintenance Coordinators (MMCo), Service Technicians (SvTn), and Machine Experts (ME). MMCos are in charge of communication and coordination of maintenance activities with customers and with SvTns. Each SvTn repairs either GM or SM machine according to his/her machine qualifications. SvTn are either shared or dedicated according to their assignment to customers. All SvTns operate from the main office and have a company car by which they drive to customer sites. They travel with all the required spare parts. MEs are

also distinguished according to their machine qualifications (GM or SM). They operate from the main office, available for phone consultation.

The traditional corrective maintenance process, described in Figure 1A, starts with a customer's fault report handled by an MMCo, who opens a service request in the CRM system, verifying the customer's service level and recording report details. Then the MMCo assigns a SvTn to the task, based on distance from customer and SvTn load, and informs the SvTn of the fault's details by email or SMS. The assigned SvTn drives to the customer's site (if time constraint allows) and analyses the fault. In case of successful analysis, the SvTn repairs the machine. If the analysis is not successful, the SvTn calls the ME for consultation. The ME analyses the fault and guides the SvTn how to repair it. After the SvTn fixes the fault, he/she updates the MMCo by SMS/ email. The MMCo closes the service request in the CRM system and checks if an additional fault has been assigned to the SvTn. If an additional fault has been assigned, the SvTn starts treating it. If not, the SvTn drives back to the main office and waits there for additional assignments.

The collaborative platform facilitates access to machine data by the maintenance service company personnel using their mobile devices. The machine data accessed includes structure, maintenance records, and state. Machine state is represented by sensor readings, e.g., temperature and pressure. The mobile access to machine data and available diagnostic apps are expected to contribute to reduce maintenance visits (single visit per fault), reduce the time it takes to repair the machine, and improve fault identification by less skilled personnel. The collaborative capabilities enable an alternative maintenance process, in which SvTns and MEs have mobile access to the customer's machine data, can perform diagnostics tasks based on machine records and sensor readings, and can remotely adjust system parameters. In addition, faults can be dynamically allocated to SvTns and there is no need for a fixed assignment when the fault report is received.

The alternative process, enabled by the ComVantage collaborative platform, is described in Figure 1B. In this process flow, customer's fault details are updated by the MMCo in the Linked Data Store (LDS), which can be viewed by all SvTns. An available SvTn chooses the next assignment using a mobile app (based on an assignment algorithm recommendation). The SvTn starts performing a remote analysis from his/her current location. In cases of access denial (e.g., in case of no internet connection or insufficient permission privileges), the SvTn contacts the MMCo who solve the problem and the SvTn can continue with the analysis. In case of successful analysis and when remote repair is possible, the SvTn repairs the fault from the remote site and marks it as repaired in the LDS. When remote repair

is not possible, the SvTn drives to the customer's site (if time constraint is satisfied) repairs the fault, marks it as repaired in the LDS, and continues to process service requests from the LDS. The MMCo finalizes the service request, marks it as repaired, and sends a message to the customer that the machine is working again. In case the problem cannot be solved by the SvTn, the fault is

allocated to an ME who can use the test results, previously performed by the SvTn, for his/her analysis. Meanwhile, the SvTn continues processing additional requests from the LDS. The ME performs the repair, if remote repair is possible. If not, the service request waits for an available SvTn.

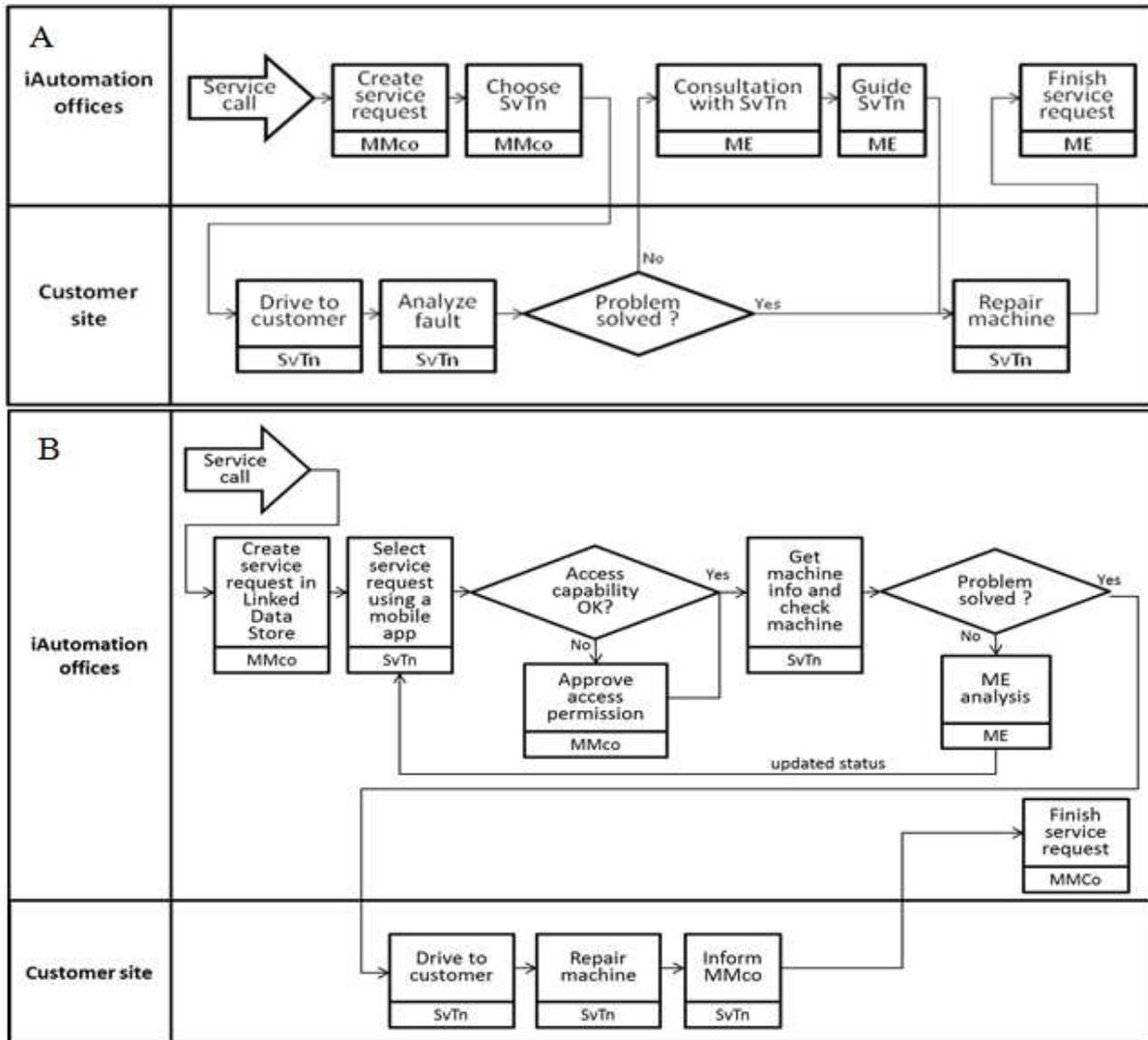


Figure 1: A) Traditional maintenance process flow, B) Collaborative maintenance process flow

SIMULATION MODELS

Model description

Simulation models, representing three process flow options, were programmed using ARENA simulation software (Rockwell Automation, USA). The “On-site Maintenance” (OM) model is for the traditional operation mode where maintenance is done only on-site and technicians are statically allocated to a fault. The other models refer to two maintenance scenarios of the ComVantage-enabled maintenance process. The “Mobile Maintenance” (MM) model is for the ComVantage enabled mobile maintenance scenario with

dynamic SvTn allocation based on fault status. The “Mobile maintenance-Permission Error” (MM-PE) model additionally incorporates the possibility of errors in permission requests.

Two average fault arrival rates were tested: a Base arrival rate and a High arrival rate. The two fault arrival rates were determined according to Mean-Time-Between-Failures (MTBF) values of each fault.. Fault inter-arrival time is exponentially distributed.

The process in the three simulation models starts with the arrival of a new service call to the MMCo and ends

once a machine is marked as repaired. In all models, machine faults occur weekdays Monday through Friday, 6:00AM to 10:00PM. During MMCo working hours, they are immediately reported. Faults that occur after MMCos working hours are reported the next morning between 8:00 and 9:00. Fault reports are treated by the MMCo sequentially according to their incoming time (full SLA have priority over partial SLA). Analysis and repair times are triangularly distributed with parameters that vary according to fault specifics.

In the MM-PE model, 10% of the permission requests incur errors requiring assistance of the MMCo. Each model was run ten times with each rate (total of 60 simulation runs). Each simulation run lasted 4 years (48 months), where half a year (six months) was regarded as warm-up time.

Analysis

Several measures were used to evaluate the implications of ComVantage-based mobile maintenance capabilities and the implication of errors in permissions requests. Mean-Time-To-Repair (MTTR) is related to quality of service and to efficiency. Average Monthly travel distance (MTD) allows assessment of sustainability and cost performance aspects. Worker utilization (ME, SvTn, and MMCo) is related to efficiency.

The experiment was designed as a repeated measure analysis of variance (ANOVA) with policy (three levels: OM, MM, MM-PE) as the within-subject factor and fault arrival rate (two levels: Base, High) as the between-subject factor. The ANOVA was followed by a Bonferroni-corrected post-hoc analysis. The Common Random Numbers (CRN) variance reduction technique was applied inducing correlation between the three models facilitating the repeated measure analysis.

RESULTS

The schedule compliance (percentage of service calls that satisfy contract time commitment) of all models was above 98%, thus indeed all models represent valid organizational operational scenarios. All main effects and interaction tested were significant at $p < 0.001$, except for ME utilization, for which the utilization did not differ between both mobile models (MM and MM-PE).

For both fault arrival rates, the mobile maintenance models (MM and MM-PE) had significantly lower MTTR than the OM model (Figure 2-Top). Indeed permission error increased MTTR in the MM-PE model relative to the MM model, yet this increase is much smaller than the difference between the MTTR for the MM and OM models. In addition, the difference between the MTTR for the different fault arrival rates in the mobile models is much smaller than in the OM case. The monthly travel distance (MTD) of the mobile

maintenance models is lower than that of the OM model (Figure 2-Bottom). The MTD for the MM-PE model is smaller than for the MM case, due to additional time required for handling the permission errors the technicians, on average service less customers in each shift, thus they drive to less sites on a daily average. This is also reflected in the prolonged MTTR in the MM-PE model relative to the MM model.

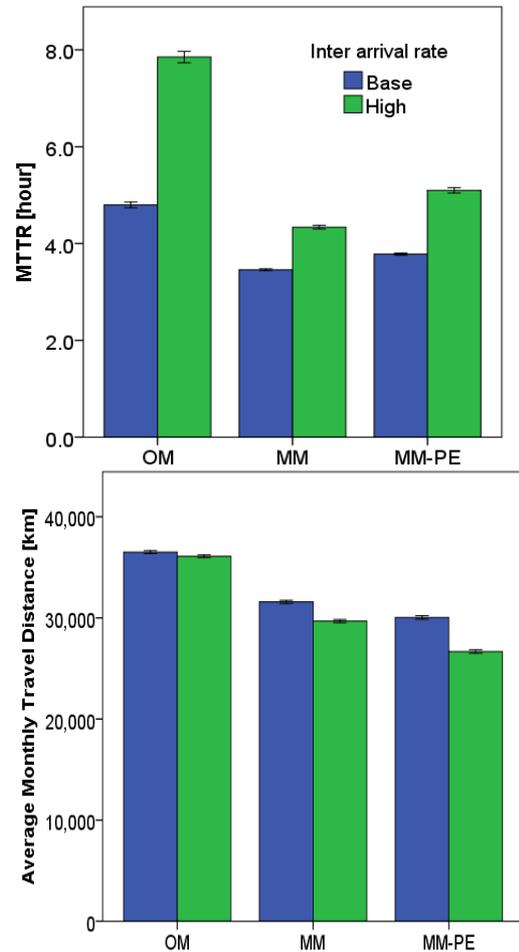


Figure 2: Top: Average monthly travel distance (MTD). Bottom: Mean time to repair (MTTR)

The MMCo utilization is similar in the different models, asserting that the collaborative system and handling permission error corrections do not overload MMCo operation (Figure 3). While average SvTn utilization in the MM model is lower than in the OM case, in the MM-PE model the SvTn utilization is the highest, which is a sign of concern regarding the effect of permission errors on SvTn workload. The utilization of the ME in both mobile models is lower than the utilization of the ME in the OM case. This asserts that not only was the ME not burdened by the collaborative system, but rather that his/her workload was reduced.

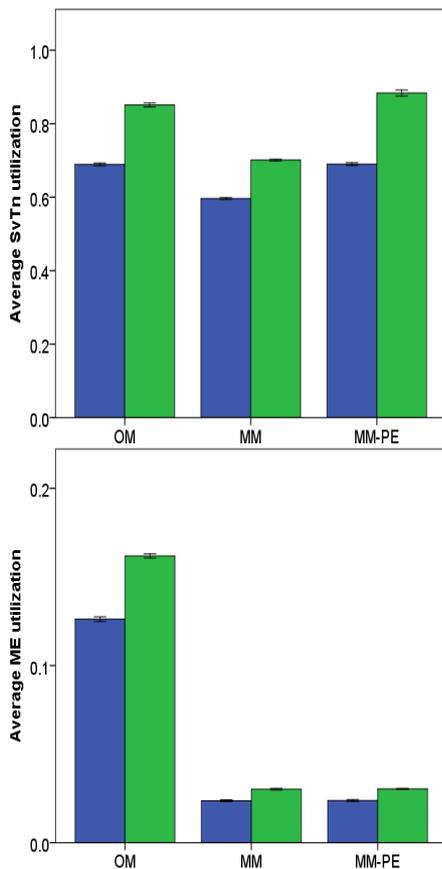


Figure 3: Top:SvTn utilization. Bottom: ME utilization

CONCLUSIONS

The results show that using the downstream capabilities provided by the collaborative platform improved the efficiency and service quality (MTTR and utilization), reduced cost, and improved sustainability (MTD). These findings are encouraging as they demonstrate a situation where improvement can be gained alongside a cost reduction. Permission errors are of concern as they negatively affect system contribution by reducing MTTR and adding to SvTn workload. This points out that investing in communication infrastructure (hardware and software) and reducing the rate of permission error events is of importance. Finally, the results show that while indeed demand volumes affect system operation, the collaborative capabilities improve system capability for effectively handling different demand volumes, improving system flexibility along the demand dimension.

This study employs the DES methodology for investigating performance impacts of emerging e-business technologies, during its development, using a process-oriented approach. Studying the impact of IT in specific business processes, at the same level at which IT is deployed, enables the research to transcend the correlational evidence between IT and business value. Therefore, this approach should complement, rather than substitute, the common, survey-based approaches. We suggest expanding the use of this approach to assess

impacts in various organizational settings. It can also be used to support decision making about which e-business technologies are effective in specific organizational circumstances.

ACKNOWLEDGMENT

The research is funded by the European Community's Seventh Framework Programme under grant agreement no. FP7-284928 ComVantage. The authors thank Dr. Werner Altmann, Frank Haferkorn, Andreas Schmid, Georg Süß, and Dr. Julianna Katona for their assistance in characterizing the case-study organization.

REFERENCES

- Dong, S., Xu, S. X., and Zhu, K. X. (2009). Research Note - Information technology in supply chains: The value of IT - enabled resources under competition. *Information Systems Research*, 20 (1), pp. 18-32.
- Fink, L. (2010). How do IT capabilities create strategic value? Toward greater integration of insights from reductionistic and holistic approaches. *European Journal of Information Systems*, 20(1), pp. 16-33.
- Lee, H. L. (2000). Creating value through supply chain integration. *Supply Chain Management Review*, 4 (4), pp. 30-36.
- Rai, A., Patnayakuni, R., and Seth, N. (2006). Firm performance impacts of digitally enabled supply chain integration capabilities. *MIS Quarterly*, 30 (2), pp. 225-246.
- Raphaelli, O., Rosenfeld, L., Fink, L., & Berman S. (2013). Assessing the Business Value of E-business: A Simulation Study of the Organizational Impacts of a Collaborative Manufacturing Network. Presented at the OASIS 2013 Pre-ICIS Workshop, Milan, Italy.
- Sanaei, Z.; Abolfazli, S.; Gani, A.; Buyya, R. 2013. "Heterogeneity in Mobile Cloud Computing: Taxonomy and Open Challenges". *IEEE Communications Surveys & Tutorials* (99): 1–24.
- Van der Vaart, T., and van Donk, D. P. (2008). A critical review of survey-based research in supply chain integration. *International Journal of Production Economics*, 111 (1), pp. 42-55.
- Zhang, X., van Donk, D. P., and van der Vaart, T. (2011). Does ICT influence supply chain management and performance?: a review of survey-based research. *International Journal of Operations & Production Management*, 31(11), pp. 1215-1247.
- Zhu. (2004). The complementarity of information technology infrastructure and e - commerce capability : a resource - based assessment of their business value. *Journal of Management Information Systems*, 21 (1), pp. 167-202.
- Zhu, K., and Kraemer, K. L. (2005). Post - adoption variations in usage and value of e-business by organizations : cross-country evidence from the retail Industry. *Information Systems Research*, 16 (1), pp. 61-84.

AUTHOR BIOGRAPHIES

ORIT RAPHAELI is a post-doctoral researcher in the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva. She received her Ph.D. in information systems from Tel Aviv University (2011) and both M.Sc and B.Sc in Industrial Engineering and Management from Ben-Gurion University. Prior to her Ph.D., she held analytical positions in the industry. Her research interests include: Business value of IT, Risk Management and Data Mining. Her e-mail address is: oritrap@post.bgu.ac.il.



LIRON ROSENFELD is a graduate student in the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva. She received her B.Sc in Industrial Engineering and Management from Ben-Gurion University. Her research interests include: Electronic markets, Mobile application market and Recommendation mechanisms. Her e-mail address is: lironr@post.bgu.ac.il.



LIOR FINK is a senior lecturer in the Department of Industrial Engineering and Management at Ben-Gurion University of the Negev. He holds a bachelor's degree in psychology and economics, a master's degree in social-industrial psychology, and a Ph.D. degree in information systems from Tel Aviv University. His research focuses on economic and strategic aspects of IT development and deployment. His e-mail address is finkl@bgu.ac.il and his Web-page can be found at <http://www.bgu.ac.il/~finkl>.



SIGAL BERMAN is a senior lecturer in the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva. She received a Ph.D. in Industrial Engineering and Management from the Ben-Gurion University (2002) and a B.Sc. in Electrical and Computer Engineering, The Technion, Haifa. Her research interests include: Human motor control, robotics and telerobotics. Her e-mail address is : sigalbe@bgu.ac.il and her Web-page can be found at <http://www.bgu.ac.il/~sigalbe/>.



STRATEGIC INFORMATION SYSTEMS PLANNING AS A DYNAMIC CAPABILITY: INSIGHTS FROM AN AGENT-BASED SIMULATION STUDY

Daniel Fürstenau
Freie Universität Berlin
School of Business & Economics
Garystr. 21, 14195 Berlin, Germany
daniel.fuerstenau@fu-berlin.de

Johannes Schinzel
Microsoft Germany
Konrad-Zuse-Str. 1, 85716 Unter-
schleißheim, Germany
johannes.schinzel@fu-berlin.de

Catherine Cleophas
RWTH Aachen
Research Area Advanced Analytics
Kackertstr. 7, 52072 Aachen, Germany
catherine.cleophas@rwth-aachen.de

KEYWORDS

Agent-based simulation, strategic IT planning, dynamic capabilities, governance and management of IT

ABSTRACT

Strategic information systems planning (SISP) helps companies to align their IT systems with their business plans. It claims to enable companies to gain competitive advantages. The resource-based view (RBV) of the firm might help to understand how and why. Drawing on prior research we believe that strategic IT planning is a dynamic capability, enabling a firm to reconfigure its resource configuration as theorized by Eisenhardt and Martin. However, this assumption has not been tested sufficiently yet. Using an agent-based simulation (ABS), this study tests to what extent strategic IT planning as a dynamic capability enables a firm to gain competitive advantages. We model an industry of companies striving for competitive advantages by optimizing their IT resources using dynamic capabilities. Given our operationalization of Eisenhardt and Martin's framework, we however cannot support the notion of SISP as a dynamic capability. Interestingly, companies in the simulation fail to realize competitive advantages because they do not anticipate competitors' moves and environmental uncertainty, an aspect deserving more attention in the resource-based view. These results and further research it may encourage demonstrate the potential of ABS for refining theories.

1 INTRODUCTION

Strategic IT planning claims to enable companies to realize competitive advantages (cf. Newkirk and Lederer 2006). By aligning the IT strategy with the corporate strategy, strategic IT planning leverages a company's IT-enabled resources. We investigate to what extent strategic IT planning acts as a dynamic capability allowing organizations to learn from previous IT investments. This sheds light on the open question how to employ IT-enabled resources to sustain competitive advantages (Tanriverdi et al. 2010). Our study is relevant for research on business and organizations as we believe that "inconclusive" evidence (Ward 2012, p. 165) on the sustained competitive advantage quest is strongly caused by methodological challenges of today's qualitative and quantitative methods. We therefore propose an alternative method of investigation for theory elaboration: An agent-based simulation.

ABS' innovative possibilities strongly influence our approach to investigate strategic IT planning as a dynamic capability: An ABS is well-equipped to formalize prior theorizing on the subject and to enable original insights on the question whether strategic IT planning acts as a dynamic capability. We propose to use ABS based on the theory building potential of simulation modeling in social science in general and in IT-related research in particular (Gilbert and Troitzsch 2011). To formulate the rules of a simulation model, the researcher has to use a set of explicit assumptions guided by theory (Axelrod 1997; Gilbert and Terna 2000; Marks 2007). This formalization process clarifies the theoretical mechanisms and boundary conditions of existing theories (Davis et al. 2007).

The contributions described in this article are theoretical as well as methodological. We link the resource-based conception of dynamic capabilities from Eisenhardt and Martin (2000) and strategic IT planning using an ABS. We move toward our theoretical objective – to underpin that strategic IT planning is a dynamic capability - by drawing on several strategic IT planning studies (refer to Ward and Peppard 2009 for general background) to formulate explicit hypotheses on the expected outcome for sustained competitive advantage. We work toward our methodological objective – to demonstrate the potential of ABS for theory building – by constructing an industry model, which is then used to test whether the simulation enables us to support the theoretical link between strategic IT planning and sustained competitive advantage. Finally, we discuss the findings of the model, theoretical implications, limitations and future research directions.

2 THEORY AND APPROACH

2.1 SISP as a Dynamic Capability

Drawing on the strategic management literature (e.g., Barney 1991; Porter 1985; Schendel 1994; Sirmon et al. 2010), our work is most closely related to dynamic capability theory from Eisenhardt and Martin (2000). The authors examine how a firm's dynamic capabilities affect its ability to achieve a sustained competitive advantage. Dynamic capabilities enable firms to alter their existing resource configuration; the more flexible resource configurations of a firm become once dynamic capabilities are present, the better the firm is able to achieve competitive advantages. We adopt Eisenhardt and Martin's (2000) definition that dynamic capabilities

are “a set of specific and identifiable processes such as product development, strategic decision making and alliancing” that “create value for firms within dynamic markets by manipulating resources into new value-creating strategies” (Eisenhardt and Martin 2000, p. 1106). Dynamic capabilities enable firms to *integrate resources*, i.e. to combine different resources to achieve a specific end, to *reconfigure resources*, i.e. to apply resource combinations successfully used in one scenario to new challenges, and to *gain and release resources*, i.e. to acquire complementary resources and shed resources that are no longer needed.

Insight into the idea of *strategic information systems planning* (SISP) as a dynamic capability is limited (Tanriverdi et al. 2010). It can be assumed that the majority of IT assets actually add value to the company so that they are not a source of competitive disadvantage (Mata et al. 1995). Owning IT assets is rather a “strategic necessity” (Powell and Dent-Micallef 1997, p. 378). If an IT asset is rare and provides an advantage, competitors will copy, acquire or substitute it (Dierickx and Cool 1989). In the case of IT assets, this usually succeeds, so that the advantage from the IT assets themselves is at best short-lived (Ross et al. 1996) and hardly the source of sustained competitive advantage. SISP as a dynamic capability, complementing IT assets, may however explain varying levels of competitive advantage among firms.

The idea that IT/IS capabilities – bundles of IT resources, competencies and practices – affect competitive advantage has attracted much research (e.g., Melville et al. 2004; Mithas et al. 2011; Mithas et al. 2012; Nevo and Wade 2010; Piccoli and Ives 2005; Wade and Hulland 2004). For instance, Bharadwaj (2000) found that some companies did not merely implement new IT, but learned lessons from past experiences and applied them in subsequent projects. Using this capability allowed them to distinguish themselves in a way that is hard to imitate and substitute. They thus achieved superior firm performance. This highlights the path dependency of IT capabilities: Companies need to develop IT capabilities and cannot copy a given capability from competitors. If a company has built IT capabilities and uses them to modify its resource configuration, this combination can prove to be immobile and consequently the source of a sustained competitive advantage and increased firm performance (Bharadwaj 2000). SISP is a process that can help companies doing so. Similarly, it was argued that IT capabilities complement IT assets and reinforce each other mutually to achieve superior firm performance (Aral and Weill 2007).

From this brief literature review, it is clear that the thrust of the existing work has been to document the impact of IT/IS capabilities on competitive advantage and performance. Unless firms do not implement SISP as a complementary learning process, however, we consider it probable that they will fail to sustain competitive advantages. Indeed, we assert that in most markets, an intermediate step – if not necessary a palliative one –

linking IT investments and firm performance is the learning process requiring strategic IT planning as a dynamic capability. Yet, despite many earlier investigations directly relating IT/IS capabilities to performance, we are not aware of empirical studies modeling the effects of SISP as a dynamic capability on competitive advantage.

2.2 Strategic IT Planning as a Dynamic Capability

One major assumption is that companies expect strategic information systems planning (SISP) to provide a competitive advantage (e.g., Galliers 1991; Kearns and Lederer 2003; Newkirk and Lederer 2006; Segars and Grover 1998). To the limited extent that existing literature demonstrates the direct performance outcomes of strategic IT planning processes, it suggests that strategic IT planning may precipitate better alignment of existing resource configurations (King and Teo 1997). There seems to be ample evidence that a systematic synchronization of a firm’s business goals and IT assets can assist in outmaneuvering competitors. For instance, Deutsche Bank currently uses its superior strategic IT planning capabilities to manage a major enterprise transformation (Gartner 2012; Deutsche Bank 2012). Costs are reduced by utilizing a core banking platform integrating both Deutsche and Postbank; in turn Deutsche is expected to become more cost efficient than its competitors in the German market. Additionally, only companies performing strategic IT planning regularly are able to adapt plans timely and sense crucial changes early on. Therefore, treating strategic IT planning as a dynamic capability should explain why regular strategic IT planning users enjoy more competitive advantages. If the assumptions of the resource-based view modeled in the simulation lead to this phenomenon, this indicates that the RBV can indeed explain this core assumption of SISP research. We therefore predict:

Hypothesis 1: If a company regularly uses SISP, it will perform better.

Since higher performance is desirable and regularly applied SISP is expected to increase performance, we can extend our first hypothesis by relating the strategic planning of a firm to the usage of SISP. According to Brews and Hunt (1999), evidence suggests that formal, strategic planning increases performance, especially in dynamic environments. Given the positive effects of SISP, companies that tend to plan ahead longer should thus be eager to use SISP to increase their competitive standing. After all, the SISP process requires substantial amounts of financial and non-financial resources, and only companies that expect long-term benefits will invest adequately (Segars and Grover 1998). When a firm plans ahead further than its competitors, it should thus be more aware of the importance of the SISP process to achieve strategic goals. Consistent with these prior lines of work, we therefore anticipate:

Hypothesis 2: If a company plans ahead further than competition, it will also use SISP techniques to a greater extent.

Although guidelines and recommendations on how to perform SISP exist (e.g., Pant and Hsu 1999; Ward and Peppard 2009), successful SISP takes more than just a step-by-step procedure. Segars and Grover (1998) and King and Teo (1997) argue that an SISP process will improve over time and will become increasingly better at integrating business and IT. For example, Premkumar and King (1994) have identified three stages of IS planning evolution. Companies more experienced in SISP should thus enjoy better performance, allowing them to realize competitive advantages at lower costs.

Hypothesis 3: If a company is more adept in SISP, it will outperform less adept competitors.

Finally, we expect that the link between SISP skills and performance is strongest in an uncertain environment. As uncertainty increases, companies should be more sensitive to their SISP capabilities and hence be more responsive to SISP capability investments, increasing the effect of SISP capability on performance. If there is little environmental uncertainty, the amount of information required for planning is lower and adapting to changes might not play an important role. In uncertain environments, however, SISP can provide advantages by helping companies to anticipate and learn about environmental dynamics (Sabherwal and Kirs 1994). Thus, companies that expertly use SISP should outperform companies with limited SISP expertise even more in uncertain environments (Newkirk and Lederer 2006).

Hypothesis 4: The higher environmental uncertainty, the greater the effect of SISP skill on performance.

3 OUR MODEL OF AN INDUSTRY

We used an ABS to create a model of an industry with competing companies. ABS allows us to control variables that cannot be influenced in reality, such as environmental uncertainty or companies' strategies.

The companies modeled as agents behave according to a set of rules, the program code. Our aim is to determine whether SISP can be considered a dynamic capability in the sense of the RBV. We thus used the assumptions of the RBV to create rules for our simulation. If our agents exhibit a behavior expected based on empirical findings of the SISP literature (similarity of simulated data and collected data, cf. Gilbert and Troitzsch 2011, p. 17), our hypotheses can be confirmed.

Modeling the regular business planning process adopted by most organizations (Teubner 2007), the simulation is turn-based. Every turn, companies choose their course of action; environmental uncertainty may alter the market, and companies having a competitive advantage will be rewarded by a score of points. Table 1 gives a high-level overview of the events occurring once at the initialization and repeatedly throughout the turns.

RBV scholars posit that to attain competitive advantage not a single resource but a specific combination of resources is necessary (e.g., Amit and Schoemaker 1993;

Barney 1991). In the simulation, this is represented by several competitive advantage optima (CAO).

Table 1. Simulation Mechanisms

Seq.	Action
I.	Randomly generate competitive advantage optima (CAO) resource configuration
II.	Randomly generate companies' resource configuration
1.*	Randomly change CAO resource configuration, depending on environmental uncertainty
2.*	Agents compare different options and select the "best" one
3.*	Agents execute the "best" option
4.*	After all agents have executed their options, the winner(s) of this round are determined
* Step repeated each turn	

As depicted in Figure 1, a CAO is a vector of randomly determined resources, e.g. (4, 7, 8). If a company is the only one that manages to fit its own resource portfolio to the CAO, it gains a competitive advantage and is rewarded with a set score. Referring to Schumpeter (1934), Barney (1991) points out that as demands and innovations shift, the resources necessary to stay ahead of competition change too. As shown in Figure 1, this environmental uncertainty is reflected in changes in the CAO's resource portfolio. Every turn, a random resource might be replaced by another, forcing the companies to adapt to stay competitive. For example, the CAO's changes from (4, 7, 8) to (4, 5, 8).

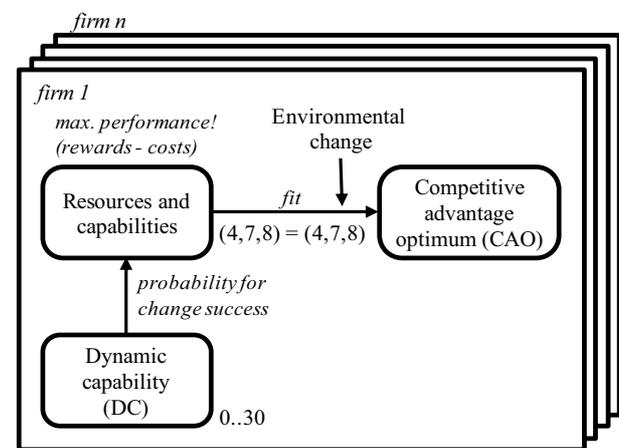


Figure 1. The model distilled

One of the most important characteristics of a dynamic capability is that it can be used to add, shed, or recombine resources (Eisenhardt and Martin 2000). In our simulation, we assume that SISP is a dynamic capability that helps companies adapt their IT and related resources to the needs of the market. Companies thus have a given SISP capability, measured on a scale from 0 to 30, and can use this capability to actively alter their resource configuration every turn, i.e. replace an existing resource with a new one. Like resources, dynamic capabilities are not distributed equally (Bharadwaj

2000). If a company is not experienced in SISP, the resource reconfiguration might go awry, and the company will not attain the desired configuration. As argued by Eisenhardt and Martin (2000), however, the company's dynamic capability will improve with exercise, even if the resource transformation is unsuccessful – in that case, the learning effect will be even greater as companies learn from small mistakes (Hayward 2002). For instance, a company might want to change its resource configuration from (4, 8, 1) to (4, 8, 5). The better the company is at a dynamic capability, the easier it is to “fit [...] the pieces together” (Powell and Dent-Micallef 1997, p. 379) and recombine its resources. In the simulation, the success of a resource change increases with the company's dynamic capability (cf. Figure 1). In our example, if the company's dynamic capability is rather low, the company will likely fail and end up with an undesirable configuration of e.g. (4, 8, 2). However, due to the experience it gained, the failure increases the firm's dynamic capability.

Not every company will try to improve its IT resource portfolio in every turn. A company might already enjoy a competitive advantage and not feel that action is necessary. As changing the resource portfolio is usually associated with costs (financial or non-financial), the company might also expect this course of action to yield less return than it incurs expenses. In this case, the company might decide to take no action during a turn. While this incurs no costs, the dynamic capability will erode as it is not exercised (Holan and Phillips 2004).

Alternatively, the company might also find that it needs to change its resource configuration, but currently does not have the necessary SISP skills to do so. In this case, the company can decide to improve its dynamic capability by spending funds on learning (Newkirk and Lederer 2006). In the simulation, the dynamic capability will increase more than by applying SISP and changing the resource configuration.

Of these three possible actions, only one can be performed in each turn. The sole goal of the agents is to maximize their overall income (i.e. the score of 100 they receive for every turn they achieve a competitive advantage) while keeping their costs (i.e. the units spent on actions – 0 for doing nothing, 5 for learning, 50 for resource changes) as low as possible, thus improving what we define as their *performance*. Companies in our model do not interact directly but are aware of each other's resource configurations.

4 EXPERIMENTAL SETUP

To test our four hypotheses, we run 24 scenarios, each with 100 runs resulting in 2.400 experiments. To set up our experiments and to determine the necessary number of runs, we thereby followed the procedure described by Law (2007, pp. 500 - 504). This section outlines the setup of input variables for each experiment performed. Table 2 gives an overview of hypotheses and the independent variables manipulated to test them.

To test H1, we used a base scenario (cf. Appendix 1). We analyzed the results of the basic setup and assigned credits for every turn a company used strategic IT planning. If a company used strategic IT planning repeatedly, the credits increased exponentially as repetitive learning facilitates success. By using this credit assignment procedure, we determined how regular a company was using its dynamic capability.

To test H2, we increased one company's propensity to plan ahead. Three experiments were performed, with this company's planning horizon (how many future turns a firm includes in its decision making process) extended from 1 to 2 and then 3 turns. Depending on how many turns ahead a company plans, it evaluates the consequences a given action (e.g. replacing a resource) has in the next turns. It also considers that it again has different choices in future turns, rendering the decision rather complex.

Table 2. Independent and dependent variable(s)

	Hypothesis	Dep.	Indep.
H1	If a company regularly uses SISP, it will perform better.	Perf.	Use of SISP
H2	If a company plans ahead longer than competition, it will also use SISP techniques to a greater extent.	Use of SISP	Planning horizon
H3	If a company is more adept in SISP, it will outperform less adept competitors.	Perf.	SISP skill
H4	The higher environmental uncertainty, the greater the effect of SISP skill on performance.	Perf.	Uncertainty in environment SISP skill

To test H3, we implemented two approaches to manipulate a company's dynamic capability. In approach 1, the company received a given initial dynamic capability, which was increased from a medium to a very high skill level over the course of five experiments. If the company decided not to apply the dynamic capability, its strategic IT planning skill would subsequently deteriorate. A high initial strategic IT planning capability was thus not guaranteed to continue throughout the simulation. In approach 2, we forced the company to use strategic IT planning in a percentage of its decisions, modeling e.g. the use of strategic IT planning to be dictated by a corporate directive. In this case we did not provide the company with a pre-determined initial strategic IT planning skill, but through subsequent application the company increased its skill. We performed three such experiments, increasing the number of cases in which the company had to choose SISP - i.e. change the resource configuration or learn - from 60% to 80% and finally 100%.

We used these alternative approaches to set up experiments for H4 as well. In this case, however, we also manipulated environmental uncertainty, i.e. the probability that a competitive advantage optima will change in a turn. Approach 1 was similar to the test of H3:

Three experiments increasing the initial SISP skill of one company from a medium to a very high level. We then performed these three experiments twice, with a 30% vs. an 80% chance of competitive advantage optima change. We used the same structure on approach 2, replicating the three experiments conducted for H3 twice, with a 30% vs. an 80% chance of competitive advantage optima change.

5 RESULTS

H1 proposed that companies using strategic IT planning more often would perform better. *Performance*, the dependent variable, is equal to the score accumulated throughout the simulation run, minus costs incurred. Using regression analysis, we found a significant link between the use of the dynamic capability and performance, but it was negative rather than positive. Companies that used strategic IT planning tended to perform worse than their counterparts. The use of strategic IT planning explained 18.8% of the difference in the dependent variable, performance.

Our experiments for H2 were evaluated using ANOVA. We analyzed the choices made by each company each turn and recoded them, with 1 standing for the use of SISP, and 2 standing for doing nothing (no SISP used). By calculating the mean of those recoded values, we can determine to what extent the companies used SISP. A mean value of 1.5 for example would indicate that the company used SISP and the “do nothing” option to a similar extent across the runs. We could confirm that companies that plan ahead longer used SISP techniques to a greater extent. However, we only found significant differences between group 1 (companies that plan ahead one turn) and groups 2 and 3 (companies that plan ahead two respectively three turns). With $p=0.495$, there was no significant difference between groups 2 and 3.

Evaluating approach 1 to test H3, we could not find a significant performance difference between those companies with an average and those with an above-average SISP skill ($p=0.411$). Approach 2 to H3 did not exhibit a significant influence of SISP skill on performance either, as our ANOVA showed ($p=0.193$): we had to reject H3 with a confidence of 95% for both approaches.

Using 2-way ANOVA, we could partially accept H4 for approach 1: As in H3, there was no significant link between SISP skill and performance as the test of between-subjects effects revealed. Neither did the covariate of the independent variables have a significant influence on the dependent variable. However, environmental uncertainty did have a significant effect on performance, in that higher environmental uncertainty led to lower performance. A 2-way ANOVA of approach 2 revealed similar results: There was a significant difference between groups with high and low environmental uncertainty, but there was no significant influence of SISP skill. The interaction effect between environmental uncertainty and SISP skill was not significant, either.

Appendix 2 includes the complete set of results for detailed analysis.

6 CONCLUDING REMARKS

We tested whether companies competing in an industry modeled after the ideas of the resource-based view of the firm achieve outcomes claimed in the strategic IT planning literature. We had to reject nearly all hypotheses. Thus we conclude that SISP did not display the characteristics of a dynamic capability in our simulation. However, we gained valuable insights into how a company’s resource configuration and its dynamic capabilities link to competitive advantage.

6.1 Limitations of the Study

The non-positive effect of a higher strategic IT planning skill on competitive advantage in our simulation might have several reasons. One reason may be that the simulation model was not correctly specified. Important aspects of the resource-based view and dynamic capabilities according to Eisenhardt and Martin (2000) might have been overlooked. We might have tilted the balance of realism and simplicity too far towards the latter in some aspects. For example, agents systematically evaluate all possible courses of actions, sometimes across multiple turns. Depending on how many resources are available, this can amount to the rational evaluation and scoring of hundreds of actions every turn, not unlike a *homo oeconomicus*. In the real world, intuition and thereby limited rationality might also influence decision making. Furthermore, there are clearly settings where the accumulation of resources is important too. We held resources abstract to focus on the influence of the dynamic capabilities, which might have been an oversimplification. However, we tried to mitigate concerns about possible deduction errors by following a transparent and comprehensible procedure to derive our hypotheses. We also did our best to set the parameters guiding the companies’ behavior to realistic levels. Still, more extensive calibration methods could increase the validity of our results.

6.2 Using ABS for Theory Building

We conclude that our example illustrates how the use of ABS helps theory building in several regards. On the one hand, creating a model to implement a simulation requires the explicit statement of a range of rules describing agents’ actions. These rules may be formulated as testable hypotheses, too, if they cannot be clearly derived from existing research. On the other hand, implementing and running a simulation can lead to surprising results. Based on the emergent properties of agent based simulations, these results cannot be foreseen before the simulation program is actually run. Once they are analyzed, reflection both on the model and on the tested hypotheses should occur. If the hypotheses tested are falsified as described here, this does not necessarily provide direct insights into the real-world problem. Rather, it shows that there appears to be no working link between the model and the expected outcome. This may

be due both to modeling assumptions and to the assumptions stated in the hypotheses. Only after careful consideration of alternative modeling and testing approaches can we consider a final rejection of the hypotheses. For this deliberate analysis, the agent based simulation has been shown to be a valuable tool.

6.3 Interpretation

We gained insights on how companies achieve and sustain competitive advantages. H1 could not be confirmed: Firms that used strategic IT planning more often performed worse. The simulation rules make sure that organizations only choose options that they expect to have a positive impact on performance. Their inability to capitalize investments may have three reasons: Their strategic IT planning capability is too low to allow them to realize the desired resource changes; due to environmental uncertainty next turn's competitive advantage situations will be different from the current to which they tried to adapt; competitors might have acquired a better or similar resource configuration.

Concerning H2 we were able to confirm that companies that plan further ahead also use strategic IT planning to greater extents. They obviously anticipate that changing their resource configuration will lead to higher performance in subsequent turns. The rejection of H1 shows that they cannot realize these expectations.

Both approaches of testing failed to support H3. Companies that are more adept in strategic IT planning (had a pronounced dynamic capability) did not outperform their competitors. Further analysis revealed that companies with high strategic IT planning skill seldom failed to improve their resource configuration. It is surprising that this improvement did not help them to gain a competitive advantage. The reasons for their inability to realize a higher performance appear to be external rather than internal. In the majority of cases, companies could not gain a competitive edge because they did not anticipate their competitors' moves or environmental uncertainty.

The experiments for H4 revealed that environmental uncertainty plays an important role in our simulated industry. If the market place changes frequently, mean performance decreases drastically.

6.4 Opportunities for Future Research

In addition to the replication and extension (the model we created is highly configurable) of our approach by peers, we see three promising ways to proceed further. First, although our model has considerable complexity, it is easy to imagine more complex interactions among agents. In particular, we find that focusing on the own resource configuration is insufficient to gain a competitive edge. The resource-based view strongly emphasizes the resources of the focal organization, following the "swings of a pendulum" (Hoskisson et al. 1999) as the currently predominant theory in strategic management after Industrial Economics. Our findings suggest that future research may benefit from considering the com-

petitive situation of a company, which is often influenced by factors outside the focal company's sphere of control. Two additional next steps may be to link the notion of strategic IT planning as a dynamic capability more tightly to models of complexity (Porter and Siggelkow 2008; Tanriverdi et al. 2010) and path dependence (Schreyögg and Kliesch-Eberl 2007) and also to increase the artificial intelligence of the agents in the simulation.

We conclude with first sketches how to extend dynamic capabilities theory in the light of our findings.

Our analysis finds that an easier reconfiguration of resources, following from higher strategic IT planning skills, was not able to explain a better performance, even in dynamic environments (i.e., refer to H3 and H4). This finding conflicts with Eisenhardt and Martin's theory. We derived it on a purely analytical level; future research might attempt to empirically clarify the boundary conditions of the relationship. Our results let us suspect that the costs for building a dynamic capability, or more specifically investments in strategic IT planning, could easily exceed performance gains. Empirically grounded data on relative cost differences between improving a dynamic capability, employing it, and not using it could help investigating this aspect further. The trade-off between investing in building a dynamic capability and the benefits from more flexibility through this dynamic capability deserves more research. How, for example, can the optimal level of investment in strategic IT planning be determined? We suspect that a more formal definition of a company's abilities to integrate, reconfigure and gain or release resources is a prerequisite for a thorough test of Eisenhardt and Martin's theory and to increase its empirical ability to predict performance differences across companies. The results of our model thus suggest that a healthy skepticism should guide future empirical studies about the extent to which strategic IT planning as a dynamic capability helps gaining a competitive advantage.

REFERENCES

- Amit, R., and Schoemaker, P. J. H. 1993. "Strategic assets and organizational rent," *Strategic Management Journal* (14:1), pp. 33–46.
- Aral, S., and Weill, P. 2007. "IT Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation," *Organization Science* (18:5), pp. 763–780.
- Axelrod, R. 1997. "Advancing the Art of Simulation in the Social Sciences," *Complexity* (Vol. (3), No. 2), pp. 21–40.
- Barney, J. 1991. "Firm Resources and Sustained Competitive Advantage," *Journal of Management* (17:1), pp. 99–120.
- Bharadwaj, A. 2000. "A resource-based perspective on information technology capability and firm performance: An empirical investigation," *MIS Quarterly* (24), pp. 169–196.
- Brews, P. J., and Hunt, M. R. 1999. "Learning to plan and planning to learn: resolving the planning school/learning school debate," *Strategic Management Journal* (20:10), pp. 889–913.

- Davis, J. P., Eisenhardt, K. M., and Bingham, C. B. 2007. "Developing theory through simulation methods," *Academy of Management Review* (32:2), pp. 480–499.
- Deutsche Bank. 2012. "Deutsche Bank launches high performance retail banking platform". Press and Media Releases. Retrieved 24th April 2013, from https://www.deutsche-bank.de/medien/en/content/3862_4172.htm
- Dierickx, I., and Cool, K. 1989. "Asset Stock Accumulation and Sustainability of Competitive Advantage," *Management Science* (35:12), pp. 1504–1511.
- Eisenhardt, K. M., and Martin, J. A. 2000. "Dynamic capabilities: what are they?" *Strategic Management Journal* (21), pp. 1105–1121.
- Galliers, R. D. 1991. "Strategic information systems planning: myths, reality and guidelines for successful implementation," *EJIS* (1:1), pp. 55–64.
- Gartner. 2012. "20 years Gartner Symposium ITxpo 2010. A Mastermind Interview With Wolfgang Gaertner, CIO of Deutsche Bank Private Banking". Retrieved 24th April 2013, from <http://gartner.mediasite.com/Mediasite/Play/c651095f6cf548c4a6033aa753c457961d>
- Gilbert, N., and Terna, P. 2000. "How to build and use agent-based models in social science," *Mind & Society* (1:1), pp. 57–72.
- Gilbert, N., and Troitzsch, K. G. 2011. *Simulation for the social scientist*, 2nd ed. [reprint], Maidenhead: Open Univ. Press.
- Hayward, M. L. A. 2002. "When do firms learn from their acquisition experience? Evidence from 1990 to 1995," *Strategic Management Journal* (23:1), pp. 21–39.
- Holan, P. M. d., and Phillips, N. 2004. "Remembrance of Things Past? The Dynamics of Organizational Forgetting," *Management Science* (50:11), pp. 1603–1613.
- Hoskisson, R. E., Hitt, M. A., Wan, W. P., and Yiu, D. 1999. "Theory and research in strategic management: Swings of a pendulum," *JoM* (25:3), pp. 417–456.
- Kearns, G. S., and Lederer, A. L. 2003. "A Resource-Based View of Strategic IT Alignment: How Knowledge Sharing Creates Competitive Advantage," *Decision Sciences* (34:1), pp. 1–29.
- King, W. R., and Teo, T. S. 1997. "Integration Between Business Planning and Information Systems Planning: Validating a Stage Hypothesis," *Decision Sciences* (28:2), pp. 279–308.
- Law, A. M. 2007. *Simulation modeling and analysis*, Boston: McGraw-Hill.
- Marks, R. E. 2007. "Validating Simulation Models: A General Framework and Four Applied Examples," *Computational Economics* (30:3), pp. 265–290.
- Mata, F. J., Fuerst, W. L., and Barney, J. B. 1995. "Information Technology and Sustained Competitive Advantage: A Resource-Based Analysis," *MIS Quarterly* (19:4), pp. p 487-505.
- Mithas, S., Ramasubbu, N., and Sambamurthy, V. 2011. "How information management capability influences firm performance," *MIS Quarterly* (35:1), pp. 237–256.
- Mithas, S., Tafti, A., Bardhan, I., and Mein Goh, J. 2012. "Information Technology and Firm Profitability: Mechanisms and Empirical Evidence," *MIS Quarterly* (36:1), pp. 205–224.
- Melville, N., Kraemer, K., and Gurbaxani, V. 2004. "Information Technology and Organizational Performance: An Integrative Model of IT Business Value," *MIS Quarterly* (28:2), pp. 283–322.
- Nevo, S., and Wade, M. R. 2010. "The formation and value of IT-enabled resources: Antecedents and consequences of synergistic relationships," *MISQ* (34:1), pp. 163–183.
- Newkirk, H. E., and Lederer, A. L. 2006. "The effectiveness of strategic information systems planning under environmental uncertainty," *Inf. & Mgt.* (43:4), pp. 481–501.
- Pant, S., and Hsu, C. 1999. *An Integrated Framework for Strategic Information Systems Planning and Development: Information Resources Management Journal (IRMJ)*: IGI Global , pp. 15–25.
- Piccoli, G., and Ives, B. 2005. "Review: It-dependent strategic initiatives and sustained competitive advantage: A review and synthesis of the literature," *MIS Quarterly* (29), pp. 747–776.
- Porter, M. E. 1985. *Competitive advantage: Creating and sustaining superior performance*, NY: Free Press.
- Porter, M., and Siggelkow, N. 2008. "Contextuality Within Activity Systems and Sustainability of Competitive Advantage," *Academy of Management Perspectives* (22:2), pp. 34–56.
- Powell, T. C., and Dent-Micallef, A. 1997. "Information technology as competitive advantage: the role of human, business, and technology resources," *Strategic Management Journal* (18:5), pp. 375–405.
- Premkumar, G., and King, W. R. 1994. "Organizational Characteristics and Information Systems Planning: An Empirical Study," *IST* (5:2), pp. 75–109.
- Ross, J. W., Beath, C. M., and Goodhue, D. L. 1996. "Develop Long-Term Competitiveness Through IT Assets," *Sloan Management Review* (38), pp. 31–42.
- Sabherwal, R., and Kirs, P. 1994. "The Alignment between Organizational Critical Success Factors and Information Technology Capability in Academic Institutions," *Decision Sciences* (25:2), pp. 301–330.
- Schendel, D. 1994. "Introduction to 'Competitive Organizational Behavior: Toward an Organizationally-Based Theory of Competitive Advantage'," *Strategic Management Journal* (15:S1), pp. 1–4.
- Schumpeter, J. A. (1934). *The theory of economic development. An inquiry into profits, capital, credit, interest, and the business cycle*. Cambridge: Harvard University Press.
- Schreyögg, G., and Kliesch-Eberl, M. 2007. "How dynamic can organizational capabilities be? Towards a dual-process model of capability dynamization," *Strategic Management Journal* (28:9), pp. 913-933.
- Segars, A. H., and Grover, V. 1998. "Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement," *MIS Quarterly* (22:2), pp. 139–163.
- Sirmon, D. G., Hitt, M. A., Arregle, J.-L., and Campbell, J. T. 2010. "The dynamic interplay of capability strengths and weaknesses: investigating the bases of temporary competitive advantage," *Strategic Management Journal* (31:13), pp. 1386–1409.
- Tanriverdi, H., Rai, A., and Venkatraman, N. 2010. "Research Commentary--Reframing the Dominant Quests of Information Systems Strategy Research for Complex Adaptive Business Systems," *Information Systems Research* (21:4), pp. 822–834.
- Teubner, R. 2007. "Strategic information systems planning: A case study from the financial services industry," *Journal of Strategic Information Systems* (16:1), pp. 105-125.
- Wade, M., and Hulland, J. 2004. "Review: The resource-based view and information systems research" *MIS Quarterly* (28), pp. 107–142.
- Ward, J. M. 2012. "Information systems strategy: Quo vadis?" *JSIS* (21:2), pp. 165–171.
- Ward, J., and Peppard, J. 2009. *Strategic planning for information systems*, Chichester: Wiley.

DANIEL FUERSTENAU is currently finishing his PhD at Freie Universität Berlin on path dependence in IT infrastructures using agent-based simulation. He studied information systems in Potsdam, Germany, and Turku, Finland. After he received a diploma degree in 2008, he worked as an IT consultant.



JOHANNES SCHINZEL was born in Munich, Germany. After finishing his BSc in Business Administration at Katholische Universität Eichstätt-Ingolstadt and Universität Antwerpen, he pursued his MSc at Freie Universität Berlin and Emory University Atlanta. In



his master's thesis he investigated the appliance of agent-based simulations in strategic IT management research. Since graduation in 2012 he works as an ERP consultant at Microsoft.

CATHERINE CLEOPHAS is Associate Professor for Advanced Analytics in the Research Area Operations Research and Management at RWTH Aachen University. She received her PhD from the University of Paderborn based on a thesis on the topic of evaluating demand forecasts for revenue management using simulation modeling in 2009. Her fields of interest include revenue management, modeling the interaction of suppliers and demand and agent-based, stochastic simulation systems.



Appendix 1 - Description of base scenario

No. of runs	100
No. of companies	10
No. of turns	30
Propensity to plan ahead (in turns)	1
All think ahead the same number of turns?	Yes
No. of distinct competitive advantages	3
No. of distinct resources	10
No. of resources in a company's resource configuration	3
No. of resources per competitive advantage optimum (CAO)	3
Maximum dynamic capability (DC)	29
Companies' choices cost restricted?	No

Maximum rent	51
Environmental uncertainty (prob. that CAO changes per turn)	30%
Probability of resource configuration change if DC max.	27%

<u>Costs for...</u>	
- changing resource configuration	50
- learning	5
- doing nothing	0

<u>Effects of..</u>	
- learning on DC	5
- doing nothing on DC	-2

<u>Effect of changing resources on DC</u>	
- if successful	1
- if failed	2

Force of SISP use c_0 in every case?	No
Force of SISP use c_0 in % of cases?	No

Appendix 2 - Descriptive statistics

Scenario	Independent variable 1	Value	Independent variable 2	Value	Dependent var.	Mean	St. dev.
1 (H1)	-	-	-	-	Performance	-109.37	546.06
2 (H2)	Propensity to plan ahead	1 turn	-	-	Chose SISP*	1.53	0.12
3 (H2)	Propensity to plan ahead	2 turns	-	-	Chose SISP*	1.36	0.17
4 (H2)	Propensity to plan ahead	3 turns	-	-	Chose SISP*	1.34	0.14
5 (H3 ¹)	Initial dynamic capability	15	-	-	Performance	-87.30	638.63
6 (H3 ¹)	Initial dynamic capability	19	-	-	Performance	-61.30	597.71
7 (H3 ¹)	Initial dynamic capability	23	-	-	Performance	16.70	578.02
8 (H3 ¹)	Initial dynamic capability	27	-	-	Performance	-18.05	609.48
9 (H3 ¹)	Initial dynamic capability	30	-	-	Performance	65.45	629.75
10 (H3 ²)	Force use of SISP in % of turns	60	-	-	Performance	-179.100	591.38
11 (H3 ²)	Force use of SISP in % of turns	80	-	-	Performance	-208.70	585.72
12 (H3 ²)	Force use of SISP in % of turns	100	-	-	Performance	-274.20	607.79
13 (H4 ¹)	Initial dynamic capability	20	Environmental uncertainty	30	Performance	-48.65	504.74
14 (H4 ¹)	Initial dynamic capability	20	Environmental uncertainty	80	Performance	-296.75	516.83
15 (H4 ¹)	Initial dynamic capability	25	Environmental uncertainty	30	Performance	4.55	588.15
16 (H4 ¹)	Initial dynamic capability	25	Environmental uncertainty	80	Performance	-353.85	342.16
17 (H4 ¹)	Initial dynamic capability	30	Environmental uncertainty	30	Performance	-43.75	345.84
18 (H4 ¹)	Initial dynamic capability	30	Environmental uncertainty	80	Performance	-352.65	292.67
19 (H4 ²)	Force use of SISP in % of turns	60	Environmental uncertainty	30	Performance	-112.70	504.89
20 (H4 ²)	Force use of SISP in % of turns	60	Environmental uncertainty	80	Performance	-234.20	640.16
21 (H4 ²)	Force use of SISP in % of turns	80	Environmental uncertainty	30	Performance	-222.80	709.63
22 (H4 ²)	Force use of SISP in % of turns	80	Environmental uncertainty	80	Performance	-400.25	276.94
23 (H4 ²)	Force use of SISP in % of turns	100	Environmental uncertainty	30	Performance	-508.45	328.56
24 (H4 ²)	Force use of SISP in % of turns	100	Environmental uncertainty	80	Performance	-636.45	371.64

*The smaller this value, the more often SISP (learning or changing resource configuration) was chosen ^{1,2}Approach 1 and 2

Extended Neonatal Metabolic Screening by Tandem Mass Spectrometry: Models and Simulation of Alternative Management Solutions

Arturo Liguori

Epidemiology and Community Medicine Unit,
Department of Paediatrics - University of Padova
Via Pietro Donà 11, 35129 Padova, Italy
e-mail: epi@pediatria.unipd.it
phone: +39-049-8215700, fax: +39-049-8215700

Giorgio Romanin-Jacur

Department of Management and Engineering
University of Padova
Stradella San Nicola 3, 36100 Vicenza, Italy
e-mail: romjac@dei.unipd.it
phone: +39-335-6072747, fax: +39-0444-998888

Abstract- Neonatal metabolic screening aims at identifying newborns with severe metabolic pathologies in order to promote appropriate interventions to avoid or to improve adverse outcomes. Tandem Mass Spectrometry permits, from a blood drop, collected on a blotting paper by a puncture on the heel, to measure a lot of metabolites according to their mass; this method can identify more than 30 metabolites, each of which is a potential marker of a hereditary metabolic disease. The large amount of available information and the difficulty in correctly interpreting them in a short time, compatible with the exigencies of newborns, imposes to find an optimal management of structures devoted to perform the related tests. In the paper four different solutions, based on different utilizations of a cluster of two or more test structures, are examined and evaluated. A simulation model, coded in language Arena, has been built to get numerical results; such a model may be usefully employed to compare the effects of different solutions for an actual situation and to give a correct dimensioning to the chosen solution.

Keywords- Metabolic hereditary diseases, neonatal screening, tandem mass spectrometry, test structure cluster, conceptual model, simulation.

1 INTRODUCTION

The screening is the application of a test to all subjects of a population with the scope of identifying a disease at the moment when it is asymptomatic. A screening test has not the significance of a diagnostic test: it identifies a person who appears to be sound but probably suffers from a disease among people who do not; people with either positive or suspicious result shall be sent to the doctor for diagnosis and necessary therapy [10, 14, 22].

The idea of an early disease diagnosis and treatment is simple; conversely the path, which brings on one side to care people for a previously undiagnosed disease and on the other side not to damage those who do not need any treatment, is not simple. Then in [37] leading criteria were fixed to identify pathologies for which a screening plan is appropriate; such criteria were repeatedly updated and are synthesized in [1, 10, 14, 16, 30].

The scope of neonatal screening lays in identifying newborns with severe pathologies in order to promote appropriate interventions to avoid or to improve adverse outcomes [35].

Biochemical mass test on newborns began in 1960 with the adoption of a screening for phenylketonuria, a rare congenital metabolism error which, if not treated, leads to a severe mental retardation. Gutrie in 1960 developed a methodology to measure phenylalanine concentration in blood; this test required a few blood drops taken from the heel and soaked into a blotting paper now known as Guthrie card [20]; moreover it has the characteristic of being quickly effected on a large sample number. Such a test first became compulsory in Massachusetts in 1963, but soon in many states neonatal screening test plans took place. Note that the same sample may be used for many tests, so that other pathologies were introduced in the neonatal screen plans.

In the 90's the development of tandem mass spectrometry brought a great change to neonatal screening, as this method permits identification of a large number of metabolites from the same sample of a few blood drops on the blotting paper, so that the screening for 30-40 metabolic pathologies is possible. Pilot plans developed in Australia and New England studied tandem mass screening effectiveness and revealed a higher identification capacity if compared with clinical diagnoses [34, 36, 38]; moreover results showed the advantage of better prognosis of identified and precociously treated patients.

Neonatal screening plans with tandem mass technology were developed in Australia, Canada, Qatar, Taiwan, Turkey and in the majority of U.S.A. [6, 11, 12, 15]. In Europe 24 states activated such plans, either applied to the whole country or limited to some regions [2, 3, 9, 13, 19, 23, 25, 26, 27]. The set of screened metabolic diseases is different among the various states.

The extended neonatal screening performed by the technique of Tandem Mass Spectrometry (MS/MS) represents an approach of absolute importance to hereditary metabolic diseases' screening [4, 5, 18, 28, 29, 32, 33]. As already mentioned above, this methods permits, from a very small volume of biological material, to measure a lot of substances of intermediate metabolism (metabolites), which form during chemical reactions transforming nutrients (like proteins and fats) into energy. Metabolites can be measured in a blood drop, collected on a blotting paper (Guthrie paper) by a puncture on the heel, by means of tandem mass spectrometry which identifies them according to their mass, which is a physical characteristic of every metabolite. This method can

identify more than 30 metabolites, each of which is a potential marker of a hereditary metabolic disease [7, 8].

MS/MS method produces large available information; on the other side such information shall be correctly and uniquely interpreted in a quick time, or anyway a time compatible with newborn exigencies (possible therapy shall be started within a time between one week and four weeks according to the specific disease); that imposes the necessity of identifying an efficient methodology for the management and organization of tests performed within neonatal screening [31]. Moreover we have to state very clearly (also on the basis of experiences reported in the literature or performed in international centres) the modes of behaviour in the case the screening yields positive results. Eventually the necessity of reducing the recall rate, i.e., the amount of newborns who are addressed to the clinical diagnosis and possible therapy, is due to the scopes of reducing both the workload of diagnosis and care structures and the stress of newborn families. Therefore it is essential to have available predictive models able to evaluate system operations in the phases of model adjustment and maintenance, related to modification of decisional trigger values for single pathologies, obtained by considering follow-ups to screening results (number of positives, false positives and negatives). A second test on the original screening sample with positive screening result, performed to specify whether the infant is affected before addressing the infant to the clinical diagnosis and possible therapy may reduce the number of false positives: such operation is known as second-tier test [17, 21]. This should be considered as part of a unique screening strategy. The versatility of MS/MS makes it possible to utilize this technology for 2-tier screening as well as for primary newborn screening, after suitable machine tuning.

Necessary equipment and structure costs, together with management and dedicated personnel costs are to be seriously considered related to expected results.

A feasible solution to many management and organizational problems is offered by the chance of using two or more machines in cluster, either with different roles or in concert to obtain a unified goal.

The paper aims at comparing different solution to utilize resources already present on the territory, in order to suggest an optimal organizational mode for a cluster of structures potentially insisting on the same catchment area. Merits and defects of every configuration are evaluated with the following objectives:

- to better manage every (positive or negative, true or false) sample follow-up, by setting a predetermined path for every possible outcome;
- to reduce the number of false negative results (and consequently the number of newborns who present pathologies but are not recognized by the system) by possibly setting an analysis path including differentiated tests characterized by high sensitivity and high specificity (2-tier);
- to rationalize the overall test method by organizing the structures so to permit, in times compatible with newborn

exigencies, the reduction of the number of newborns addressed to clinical diagnosis;

- to reduce times necessary to taken samples to cross the whole diagnostic system;
- rationalize equipments and structures utilization by considering the overall number of treated samples;
- rationalize territorial resources and teams utilization so to avoid useless overlapping and to obtain an effective and unique coordination.

Moreover the paper presents both conceptual and simulation models able to compare different distributions of resources and tasks to two or more centres serving the same catchment area.

2 BASIC ASSUMPTIONS

While configuring the suggested models some basic elements of the literature were considered.

2.1 Numerical dimension of catchment area

The correct operation of a screening system depends on the knowledge of markers within people who does not suffer from diseases (reference values) and within people who suffer from diseases, on the correct setting (and successive tuning on the basis of feedback analysis) of the Cut-off to separate the two populations, on the definition of precise decision trigger values for any consequent action.

The applied solution shall permit to compute reference values on the basis of a sufficiently large population, by using the same diagnostic instrument, so to compare real results detected on the territory with reference values reported in the literature, with the scope of increasing the quality of extended neonatal screening).

In the configuration of many stand alone centres the typical dimension of 30,000 samples per year is not reached, while on the basis of international experiences reported in the literature a minimum number of 35,000 and an optimal one of at least 50,000 samples per year is suggested, justified by the following elements:

- plant and structure cost amortization (the break even is normally placed at 35,000 samples per year);
- rarity of some pathologies does not permit, particularly with small catchment area, to build up a sufficient experience for operators and above all to correctly set trigger values and cut off points for the population where the centre is placed, on the basis of analytical results after the screening phase;
- suboptimal dimensions of catchment areas of many projects make the screening system self-learning period particularly long and difficult [31].

Note that the dimension of catchment area for a single centre may reach values much larger than 50,000 samples per year, for instance in Europe in 2007 there were centres able to manage up to 77,000 samples per year [3].

2.2 Structure specialization

The configuration of an only screening cluster based on the specialization of two or more structures within it actually

permits to obtain a remarkable improvement in the quality of performed analyses. That mainly happens as, if correctly dimensioned and managed, a cluster can permit either a different setting of system machines, with no risk of increasing the number of false negatives, or the use of the same machine with frequent setting changes between high sensitivity and high specificity, without substantially affecting the sample crossing time.

3 COMPARISON AMONG CONCEPTUAL MODELS

3.1 Generic system

A generic system which would be applied in the case of an only centre, considered only as a reference, is represented by a block diagram in Figure 1, where the sample movement inside the system is clearly reported.

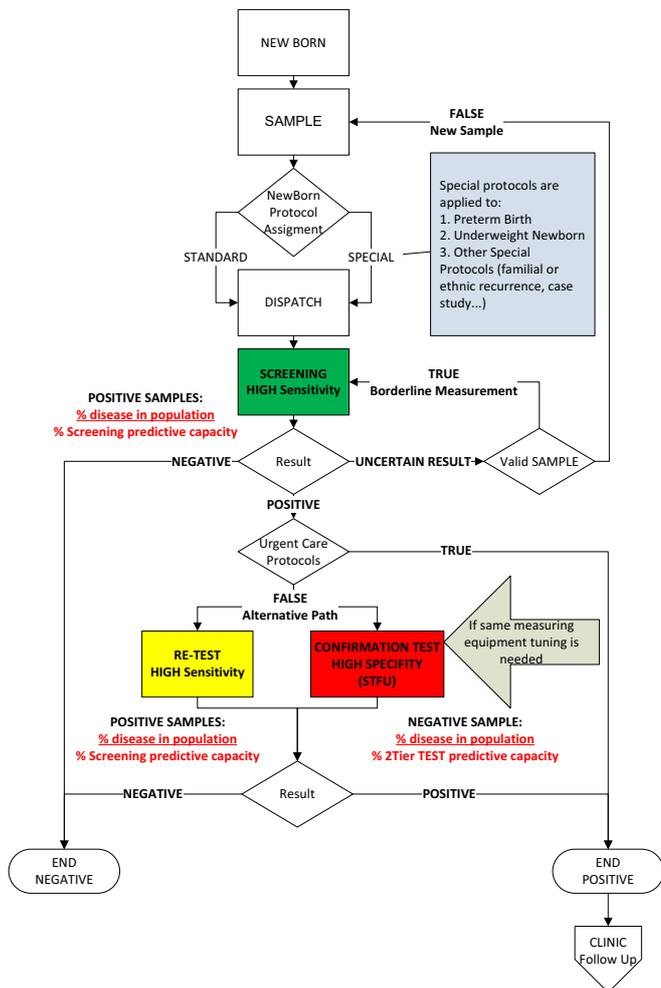


Figure 1. Generic test system.

The presented model describes the standard operation of a screening centre performing sample test and possible retest in case of positive result, either at high sensitivity or at high specificity (second tier) after a new machine tuning; the second confirmation test at high specificity is also called Short Term Follow Up (STFU).

The model can be simulated by using three parameters for every pathology:

- the pathology incidence on the population
- the predictive capacity of the specific test (true positives/true+false positives) observed in the screening phase, with measurement in condition of high sensitivity
- the predictive capacity of the specific test (true positives/true+false positives) observed in the screening phase, with measurement in condition of first tier high sensitivity test and second tier high specificity test, after different machine tuning.

The four studied systems with two analysis centres are represented by block diagrams in Figures 2, 3, 4, 5, which represent four different cluster configurations of screening structures:

1. Cluster with geographical distribution of population on the specific competent centre, where structures perform only high sensitivity screening;
2. Cluster with geographical distribution of population on the specific competent centre, where one structure performs only high sensitivity screening and the other not only performs high sensitivity screening but is at disposition for second tier analysis at high specificity with different machine tuning;
3. Cluster with geographical distribution of population on the specific competent centre, where both structures perform not only high sensitivity screening but also second tier analysis at high specificity with different machine tuning;
4. Cluster with centres specialization and differentiation in terms of catchment areas and utilization of machines and structures.

3.2 Cluster with geographical population distribution, only high sensitivity screening

The first solution described in Figure 2 represents the most common current configuration: two or more centres, even close to one another for what concerns the catchment area, are utilized by simply allocating the population according to territorial competence, trying at most to get numerical balance. In this case we hypothesize that centres limit their activity to performing a high sensitivity analysis so to reduce at most the amount of false negatives and in the meantime to keep the sample crossing time restrained. In any case the retest is planned for those samples which in the first instance give positive result, but without adopting different tuning for the screening machine (in order to reduce the amount of false positives).

The advantages are:

- restraint of false negatives amount;
- high sample screening speed;

and the disadvantages are:

- high amount of false positives (and consequent address to clinical structure for diagnosis, possible therapy and follow-up planning), not restrained by a high specificity second tier test able to increase predictive capacity.

This method proves to be especially effective in the case the catchment area related to every centre is close to maximum analysis capacity of available machines and structures.

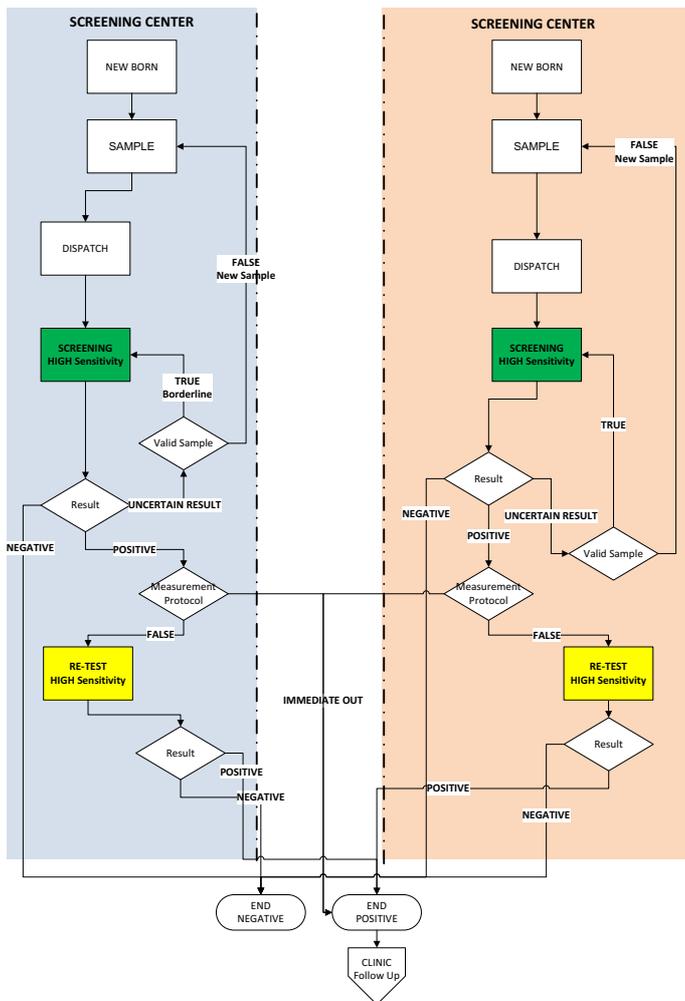


Figure 2. Cluster with geographical population distribution, only high sensitivity screening.

The block diagram in Figure 2 is divided into two parts which represent both the geographical placing and the different responsibilities of involved institutions.

The model operation is the following:

1. The sample is taken and sent to the competent centre;
2. A first high sensitivity screening is performed (to reduce the chance of a false negative);
3. In case of uncertain result the test is newly performed if the sample is good, otherwise a new sample is taken and the test repeated;
4. If the screening gives negative outcome, the sample goes out of the system, otherwise a retest is performed on the same machine with the same tuning;
5. In case of confirmation of positive outcome by the retest the newborn is addressed to the clinical centre for diagnosis, possible therapy and follow up.

3.3 Cluster with geographical distribution of population on the specific competent centre, where one structure performs high sensitivity screening and the other also performs high specificity retest.

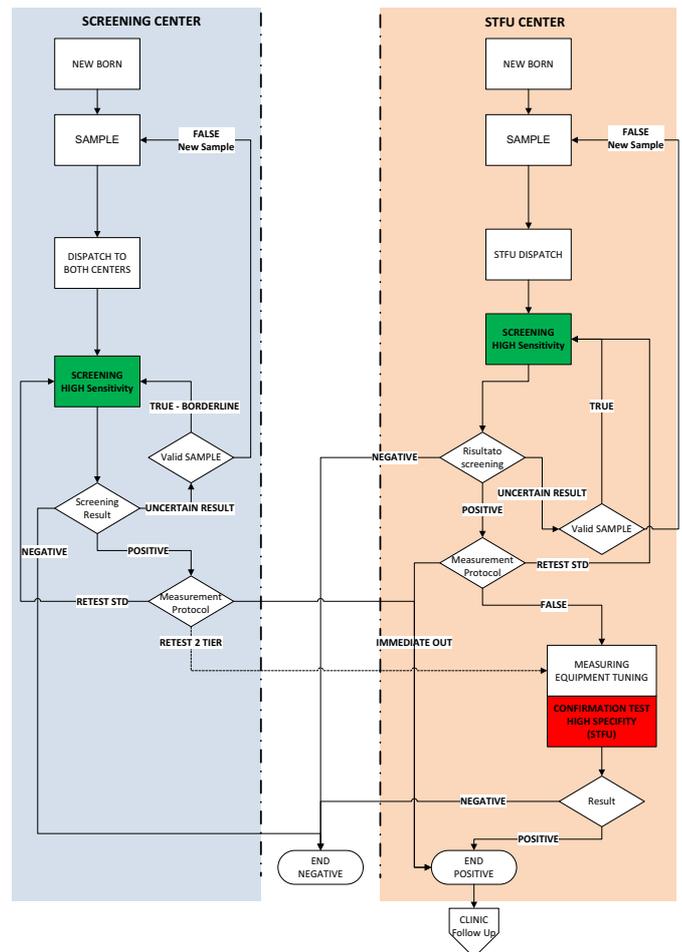


Figure 3. One structure performs high sensitivity screening, the other performs high specificity retest too.

The second solution described in Figure 3 represents the configuration where one of the two centres available in the considered geographical area is able to perform both high sensitivity and high specificity test. This is the typical case where to a centre with structures and personnel able to perform 2-tier screening a new centre able to perform only high sensitivity screening for its own area is associated. This configuration is frequently identified in the literature by remarking that one of the centres either “performs the Short Term Follow Up function”, or “has the ability of performing high specificity analysis”, and consequently has a better predictive ability.

Note that if the second tier centre has only one machine at disposition for analyses, such a machine may be utilized for both high sensitivity and high specificity analysis, therefore it is necessary to consider the new tuning activity to get a different machine setting; such an activity may be effected either when necessary or periodically (for instance once a week) with different results for what concerns the overall sample crossing time (both for negative and positive ones). Another factor affecting the mean sample crossing time is the necessity of transferring the sample to the centre able to perform high specificity analysis whenever the high

sensitivity analysis has given positive result in the other centre: to face the related difficulties it is generally suggested to send the samples simultaneously to both centres, when they are attributable to the centre performing only high sensitivity analysis.

A final notation about the model concerns the adoption of an “immediate output”, with exit towards the clinical diagnosis and possible therapy even in absence of MS/MS retest, in case the sample results to be positive with respect to some pathologies associated with very rapid clinical evolution.

Advantages presented by this model are the following:

- The amount of false positives is reduced;
- It is possible to manage the cluster in a centralized and coordinated way (for instance for trigger setting);
- It is possible to specialize teams and organization in a different way in the two centres.

The only disadvantage is the increasing of sample crossing time.

This method is valid if in the catchment area where the two centres are competent a team with greater experience in the management of metabolic diseases is present, to which the whole cluster is entrusted.

3.4 Cluster with geographical distribution of population on the specific competent centre, where both structures perform both high sensitivity screening and high specificity retest when applicable.

The third solution described in Figure 4 represents the configuration where both centres independently effect the whole test-retest path with different setting on the catchment area where they are competent. Actually it represents a system of independent centres utilizing a method which permits to reduce the amount of false positives addressed to the successive phase of clinical diagnosis. The hypothesis of considering the two centres pertaining to an only cluster is represented by the fact that all decision elements (for instance, trigger values) are ruled by an only direction.

The advantage with respect to the previous model is given by a substantial reduction of sample crossing times.

The disadvantages are the following:

- In order to operate correctly the structures shall work for a greater amount of time;
- In both structure special personnel, able to effect machine tuning and with greater experience in reading screening results, is necessary.

3.5 Centres with differentiated purposes

The difference of this model, described in Figure 5, with respect to the previous ones stays in the fact that centres perform activities which are only partially superimposable, as they are devoted to develop different activities and specializations, related to a high sensitivity and a high specificity structure.

Samples are allocated to centres not on the basis of geographical origin but according to different criteria; in the

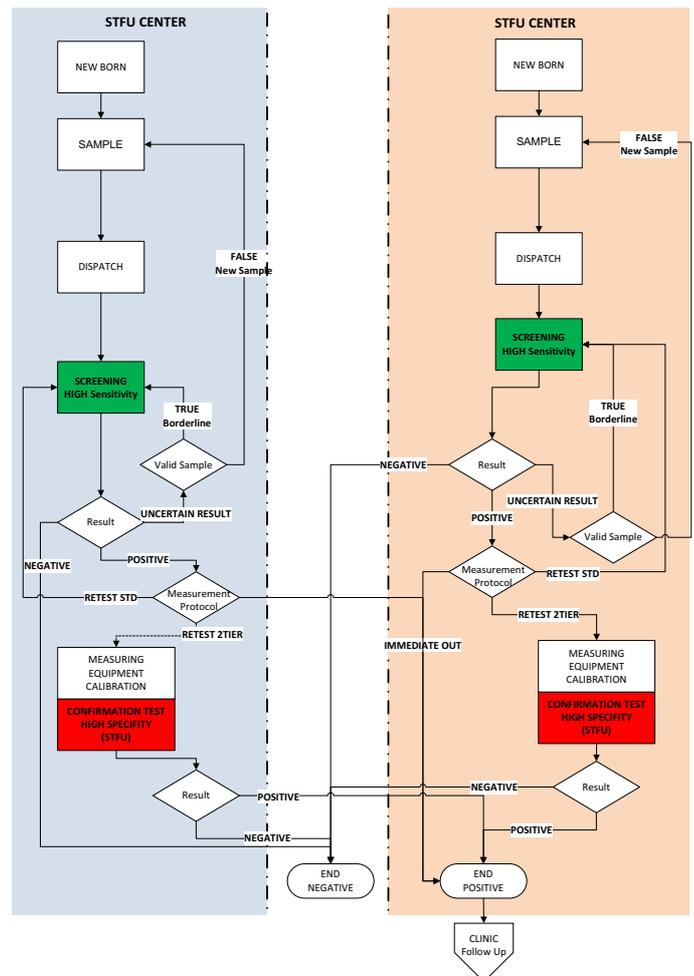


Figure 4. Both structures perform high sensitivity screening and high specificity retest.

model hypothesis all samples are simultaneously sent to the two centres (double sample and double sending) except for those which belong to the “special protocol” category (second samples in case of uncertain result, preterm and underweight infants, familial or ethnic recurrences, etc.) which are directly sent to the high specificity centre.

The principal benefit of this approach lays in permitting to both centres structures to build up a sufficient experience to reach a high qualitative standard. In fact:

- the high sensitivity centre gets a very large amount of samples, corresponding to a larger catchment area with respect to previous models; anyway that is manageable because this centre has to operate only high sensitivity screening and to communicate to the high specificity centre the reference of samples on which the retest is to be effected (possibly with different machine tuning); the high sensitivity centre operation is effected without off-line or non serialized activities;
- the high specificity centre, devoted to the analysis of samples requiring retest and to protocols requiring non standard treatments, on the contrary has a much smaller amount of samples to analyze; therefore it can examine a

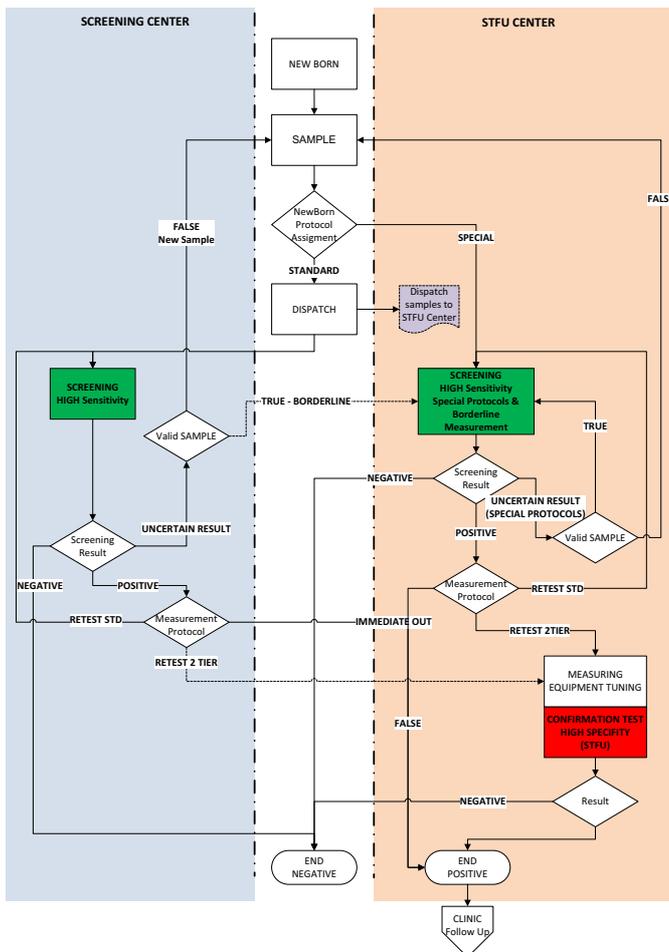


Figure 5. Centres with differentiated purposes.

large amount of important samples following ad hoc procedures without an excessive increase of crossing time, as they are managed by a machine able to work much smaller lots, sometimes of only one sample.

Therefore this model, differently from other suggested configurations, permits to utilize different settings for the machines of the various structures, also in a permanent way and without requiring any tuning. Then the screening centre may be set with a configuration devoted to false negatives reduction (high sensitivity centre); the false positives amount will be obviously larger with respect to what is commonly given in the literature; the hypothesized structure is sufficiently dimensioned also with a comparatively large birth catchment area (100,000 units per year). On the contrary the second structure of the cluster system may be adapted to a greater specialization, as it has to manage a smaller sample amount, and such samples are always important, as either they come from newborns belonging to particular groups or they have been indicated as positive by a first screening.

This double structure (screening and confirmation) with contextual availability of blood samples has the advantage of effecting off-line all tests necessary to reduce the chance of false positives without stopping the routine operation of the

other centre (with consequent machine tuning); such a way useless addresses to clinical diagnosis are avoided, with consequent benefit to the parents' mental health.

The main advantages are the following:

- This configuration permits an optimal sample distribution in terms of analysis quality (building of personnel experience, territorial personnel distribution so to avoid duplicates, specialization of personnel devoted to positivity confirmation);

- This model is easily scalable as it permits to dimension the catchment area with respect to one or more high sensitivity screening centres with a reference centre for high specificity analyses;

- With respect to previous models a substantial reduction of positive sample crossing times is reached, above all in the hypothesis of having permanently machines with differentiated tuning.

The main disadvantage is due to costs of double withdrawal and double sending.

4 SIMULATION MODEL

In order to reach the scopes indicated in the introduction we built up a simulation model, coded in language Rockwell Arena, able to describe simultaneously the four presented conceptual models.

4.1 Macroscopic description

Here we present in Figure 6 a macroscopic description of the simulation model by means of submodels; from left to right we see:

1. Sample input in the system (coinciding with births);
2. Determination of positive samples (successively detected) and of special protocols;
3. Sample dispatch to the competent centre (after sample doubling in order to let the four different models operate simultaneously);
4. Models corresponding to the four conceptual solutions seen in Section 3.

Sample inputs are scheduled according to a Poisson process. Consider now what represented in Figure 7. In the considered conceptual models a sample, after withdrawal and transport, proceeds to the test and get either positive or negative result, as shown to the left in the figure. In simulation, in order to increase clarity and data manageability, the procedure shown to the right in the figure is adopted; to every sample either a pathology from the panel of considered ones or no pathology is probabilistically allocated, i.e., either a positive attribute to one disease or no attribute is assigned, and consequently the probability of all possible results is given, both related to high sensitivity screening only and related to high sensitivity screening plus high specificity confirmation; such a way the test result is deterministically obtained later, by knowing whether the single or the double test has been applied.

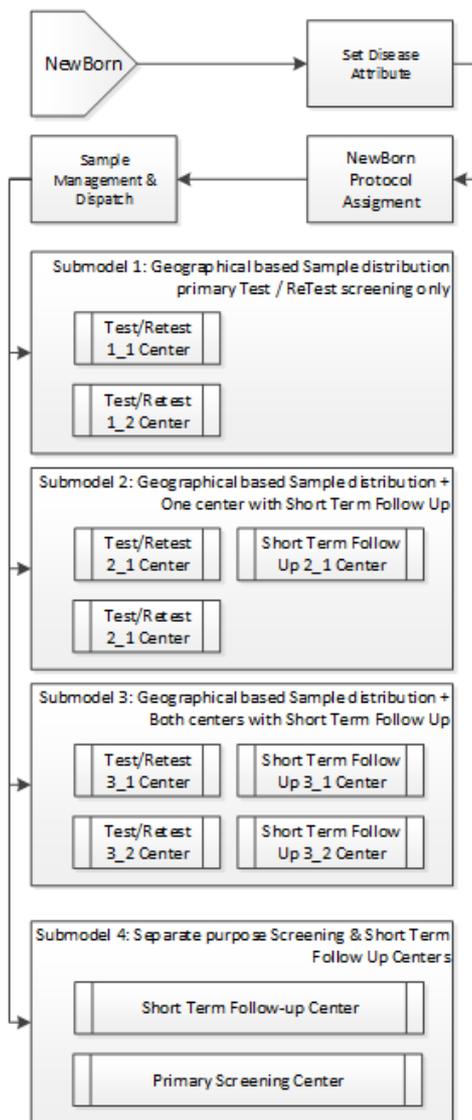


Figure 6. Simulation model macroscopic description

By a probabilistic multiple decision all non standard protocols (preterm, underweight, special) are defined by a further attribute.

All above probabilities are obtained either from the literature or from direct experience on the actual situation.

Sample transportation by an actual carrier with related time constraints is simulated.

Finally the four alternative models, describing the above mentioned alternative configurations, are coded in Arena, reproducing the corresponding conceptual models with actual operation parameters. The whole Arena model is reported in Figure 8.

4.2 Simulation results

The simulation results, related to an actual situation and to every cluster type, include:

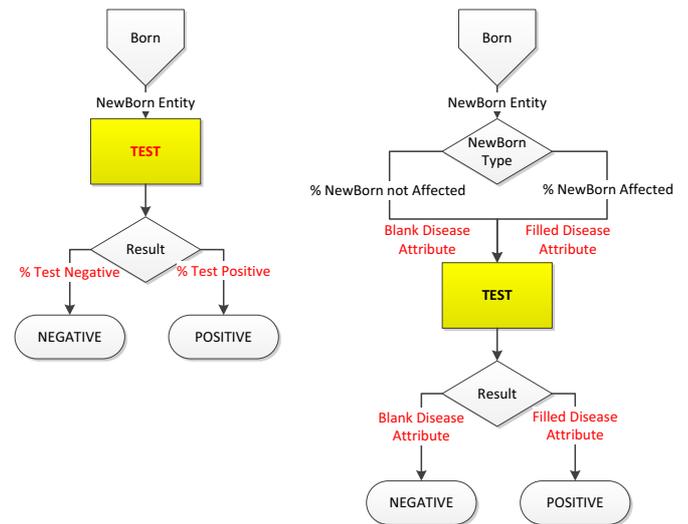


Figure 7. Conceptual and simulation oriented sample parametrization.

- Crossing time: for every pathology it is checked whether the crossing time is acceptable with respect to the risk of delays in the diagnosis and therapy phase with consequent permanent damages to positive newborns;

- Workload: both workload and performance of test structures (machines, personnel, etc.) and workload of clinical structures devoted to diagnosis and therapy are computed.

The above considered results may be usefully studied to compare possible solutions and to suggest a correct dimensioning for the chosen configuration.

4.2 Simulation model applications

We have to remark that data from the literature about metabolic hereditary diseases incidence and related detectability, based on screenings performed in various countries, are notably different from one another. This is due to two different causes: the first is connected to population characteristics, the second to cut-off values adopted to distinguish positive and negative patients from tandem mass spectrometry results. Therefore the screening process can be optimized only after a sufficiently large time interval (generally at least two years), to know incidence parameters with good precision. In the meantime we may only base our actions on a value interval, between minimum and maximum values taken from the literature.

The model has been used to rightly dimension the screening process in Veneto Region (North-East Italy), where two screening centres are set. In correspondence with minimum and maximum incidence values from the literature, and for different cut-off values, the amount of positive patients both non detected and non treated in time has been evidenced by running the proposed model; as a consequence the operating rules for both screening centres (working time for every day of the week) and clinical follow-up centres (number of patient individual visits and investigations) have been

obtained to minimize the probability of damaged patients. As a first application, the first configuration (cluster with geographical population distribution, only high sensitivity screening), was considered. As useful alternatives, all other configurations, with related results corresponding to different adopted cut-off values, were presented and will be taken into consideration by regional health care managing authorities. In the meantime, local results will permit to refine population based characteristics in term of diseases incidence.

REFERENCES

- [1] Andermann A, Blancquaert I, Beauchamp S, Déry V. (2008) Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. *Bull World Health Organ.* 2008 Apr; 86(4): 317-9.
- [2] Baehner F, Schmiedeskamp C, Krummenauer F, Miebach E, Bajbouj M, Whybra C, Kohlschütter A, Kampmann C, Beck M. (2005) Cumulative incidence rates of the mucopolysaccharidoses in Germany. *J Inherit Metab Dis.*; 28(6): 1011-7.
- [3] Bodamer OA, Hoffmann GF, Lindner M. (2007) Expanded newborn screening in Europe 2007. *J Inherit Metab Dis.*; 30(4): 439-44. Epub 2007 Jul 23.
- [4] Burton H, Moorthie S. (2010) Expanded newborn screening. A review of the evidence, NHS, NIH CLAHRC, PHG foundation, May 2010.
- [5] Chace DH, Kalas TA, Naylor EW. (2002) The application of tandem mass spectrometry to neonatal screening for inherited disorders of intermediary metabolism. *Annu Rev Genomics Hum Genet.*
- [6] Chien YH, Chiang SC, Zhang XK, Keutzer J, Lee NC, Huang AC, Chen CA, Wu MH, Huang PH, Tsai FJ, Chen YT, Hwu WL. (2008) Early detection of Pompe disease by newborn screening is feasible: results from the Taiwan screening program. *Pediatrics.*; 122(1):e39-45. Epub 2008 Jun 2.
- [7] Dajnoki A, Mühl A, Fekete G, Keutzer J, Orsini J, Dejesus V, Zhang XK, Bodamer OA. (2008) Newborn screening for Pompe disease by measuring acid alpha-glucosidase activity using tandem mass spectrometry. *Clin Chem.*; 54(10):1624-9. Epub 2008 Aug 14.
- [8] Dajnoki A, Fekete G, Keutzer J, Orsini JJ, De Jesus VR, Chien YH, Hwu WL, Lukacs Z, Mühl A, Zhang XK, Bodamer O. (2010) Newborn screening for Fabry disease by measuring GLA activity using tandem mass spectrometry. *Clin Chim Acta.* 411(19-20): 1428-31. Epub 2010 Mar 22.
- [9] Dionisi-Vici C, Rizzo C, Burlina AB, Caruso U, Sabetta G, Uziel G, Abeni D. (2002) Inborn errors of metabolism in the Italian pediatric population: a national retrospective survey. *J Pediatr.*; 140(3): 321-7.
- [10] Fernandes J, Saudubray JM, van der Berghe G, Walter JH. (2006) *Inborn Metabolic Diseases. Diagnosis and Treatment.* Fourth, revised edition. Springer.
- [11] Feuchtbaum L, Lorey F, Faulkner L, Sherwin J, Currier R, Bhandal A, Cunningham G. (2006) California's experience implementing a pilot newborn supplemental screening program using tandem mass spectrometry. *Pediatrics.*; 117(5 Pt 2): S261-9.
- [12] Frazier DM, Millington DS, McCandless SE, Koeberl DD, Weavil SD, Chaing SH, Muenzer J. (2006) The tandem mass spectrometry newborn screening experience in North Carolina: 1997-2005. *J Inherit Metab Dis.*; 29(1): 76-85.
- [13] Hoffmann GF, von Kries R, Klose D, Lindner M, Schulze A, Muntau AC, Röschinger W, Liebl B, Mayatepek E, Roscher AA. (2004) Frequencies of inherited organic acidurias and disorders of mitochondrial fatty acid transport and oxidation in Germany, *Eur J Pediatr.*; 163(2):76-80
- [14] Holtzman C, Slazyk WE, Cordero JF, Harmon WH. (1986) Descriptive epidemiology of missed cases of phenylketonuria and congenital hypothyroidism. *Pediatrics.*; 78:553-558
- [15] Hwu WL, Chien YH, Lee NC, Chiang SC, Dobrovolny R, Huang AC, Yeh HY, Chao MC, Lin SJ, Kitagawa T, Desnick RJ, Hsu LW. (2009) Newborn screening for Fabry disease in Taiwan reveals a high incidence of the later-onset GLA mutation c.936+919G>A (IVS4+919G>A). *Hum Mutat. Oct.*; 30(10): 1397-405.
- [16] Kemper AR, Hwu WL, Lloyd-Puryear M, Kishnani PS. (2007) Newborn screening for Pompe disease: synthesis of the evidence and development of screening recommendations. *Pediatrics.* 120(5): e1327-34.
- [17] La Marca G, Malvagia S, Casetta B, Pasquini E, Donati MA, Zammarchi E. (2008) Progress in expanded newborn screening for metabolic conditions by LC-MS/MS in Tuscany: Update on methods to reduce false tests. *J Inherit Metab Dis.* 2008 Oct 27.
- [18] Lindner M, Hoffmann GF, Matern D. (2010) Newborn screening for disorders of fatty-acid oxidation: experience and recommendations from an expert meeting. *J Inherit Metab Dis.* 2010 Apr. 7.
- [19] Maier EM, Krone N, Busch U, et al (2002) Medium chain acyl-CoA dehydrogenase (MCAD) mutations in patients identified by prospective MS/MS-based newborn screening in Bavaria. Abstracts book of the 5th meeting of the International Society for Neonatal Screening. Genova 2002.
- [20] Marsden D, Levy H. (2010) Newborn screening of lysosomal storage disorders. *Clin Chem.*; 56(7): 1071-9. Epub 2010 May 20. Review.
- [21] Marsden D., Larson C., Levy H. (2006) Newborn screening for metabolic disorders, *Journal of Pediatrics.*; 148(5): 577-584
- [22] Minichiello C. (2010) *Farmaci e malattie rare. Programmazione delle reti per il trattamento dei malati di una ampia area del nord Italia.* PhD thesis – University of Padova, Italy.
- [23] Moore D, Connock MJ, Wraith E, Lavery C. (2008) The prevalence of and survival in Mucopolysaccharidosis I: Hurler, Hurler-Scheie and Scheie syndromes in the UK. *Orphanet J Rare Dis.*; 3:24.
- [24] Pampols T. (2003) Neonatal screening. *The Turkish Journal of Pediatrics.*; 45: 87-94
- [25] Poorthuis BJ, Wevers RA, Kleijer WJ, Groener JE, de Jong JG, van Weely S, Niezen-Koning KE, van Diggelen OP

(1999) The frequency of lysosomal storage diseases in The Netherlands. *Hum Genet.*; 105(1-2): 151-6.

[26] Poupetová H, Ledvinová J, Berná L, Dvoráková L, Kozich V, Elleder M. (2010) The birth prevalence of lysosomal storage disorders in the Czech Republic: comparison with data in different populations. *J Inher Metab Dis.*; 33(4): 387-96.

[27] Sanjurjo P, Baldellou A, Aldámiz-Echevarría K, Montejo M, García Jiménez M. (2008) Inborn errors of metabolism as rare diseases with a specific global situation. *An Sist Sanit Navar.*; 31 Suppl 2: 55-73. Spanish.

[28] Schulze A, Lindner M, Kohlmüller D, Olgemöller K, Mayatepek E, Hoffmann GF. (2003) Expanded newborn screening for inborn errors of metabolism by electrospray ionization-tandem mass spectrometry: results, outcome, and implications. *Pediatrics*; 111(6 Pt 1): 1399-406.

[29] Schweitzer S. (1995) Newborn mass screening for galactosemia, *European Journal of Pediatrics* ; 154(7 Suppl 2): S37-9

[30] Scriver CR, Beaudet AL, Sly WS, Valle D. Childs B, Kinzler KW, Vogelstein B. (2001) *The Metabolic and Molecular Bases of Inherited Disease*, Eighth Edition McGraw-Hill.

[31] SISMME-SISN Società Italiana Studio Malattie Metaboliche Erediatrie – Società Italiana Screenings Neonatali. (2008) *Linee guida per lo screening neonatale esteso e la conferma diagnostica.*

[32] Spada M, Pagliardini S, Yasuda M, Tukul T, Thiagarajan G, Sakuraba H, Ponzone A, Desnick RJ. (2006) High incidence of later-onset fabry disease revealed by newborn screening. *Am J Hum Genet.*; 79(1): 31-40.

[33] Vilarinho L, Rocha H, Sousa C, Marcão A, Fonseca H, Bogas M, Osório RV. (2010) Four years of expanded newborn screening in Portugal with tandem mass spectrometry. *J Inher Metab Dis.*; 33 Suppl 3: S133-8.

[34] Wilcken B, Wiley V, Hammond J, Carpenter K. (2003) Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N Engl J Med.* 2003 Jun 5; 348(23): 2304-12.

[35] Wilcken B, Wiley V. (2008) Newborn screening. *Pathology.*;40(2):104-15. Review.

[36] Wiley V, Carpenter K, Wilcken B. (1999) Newborn screening with tandem mass spectrometry: 12 months' experience in NSW Australia. *Acta Paediatr Suppl.* 1999 Dec; 88(432): 48-51.

[37] Wilson J. M., Jungner G. (1968). "Principles and practice of screening for disease." World Health Organization .

[38] Zytkovicz TH, Fitzgerald EF, Marsden D, Larson CA, Shih VE, Johnson DM, Strauss AW, Comeau AM, Eaton RB, Grady GF. (2001) Tandem mass spectrometric analysis for amino, organic, and fatty acid disorders in newborn dried blood spots: a two-year summary from the New England Newborn Screening Program. *Clin Chem.*; 47(11): 1945-55.

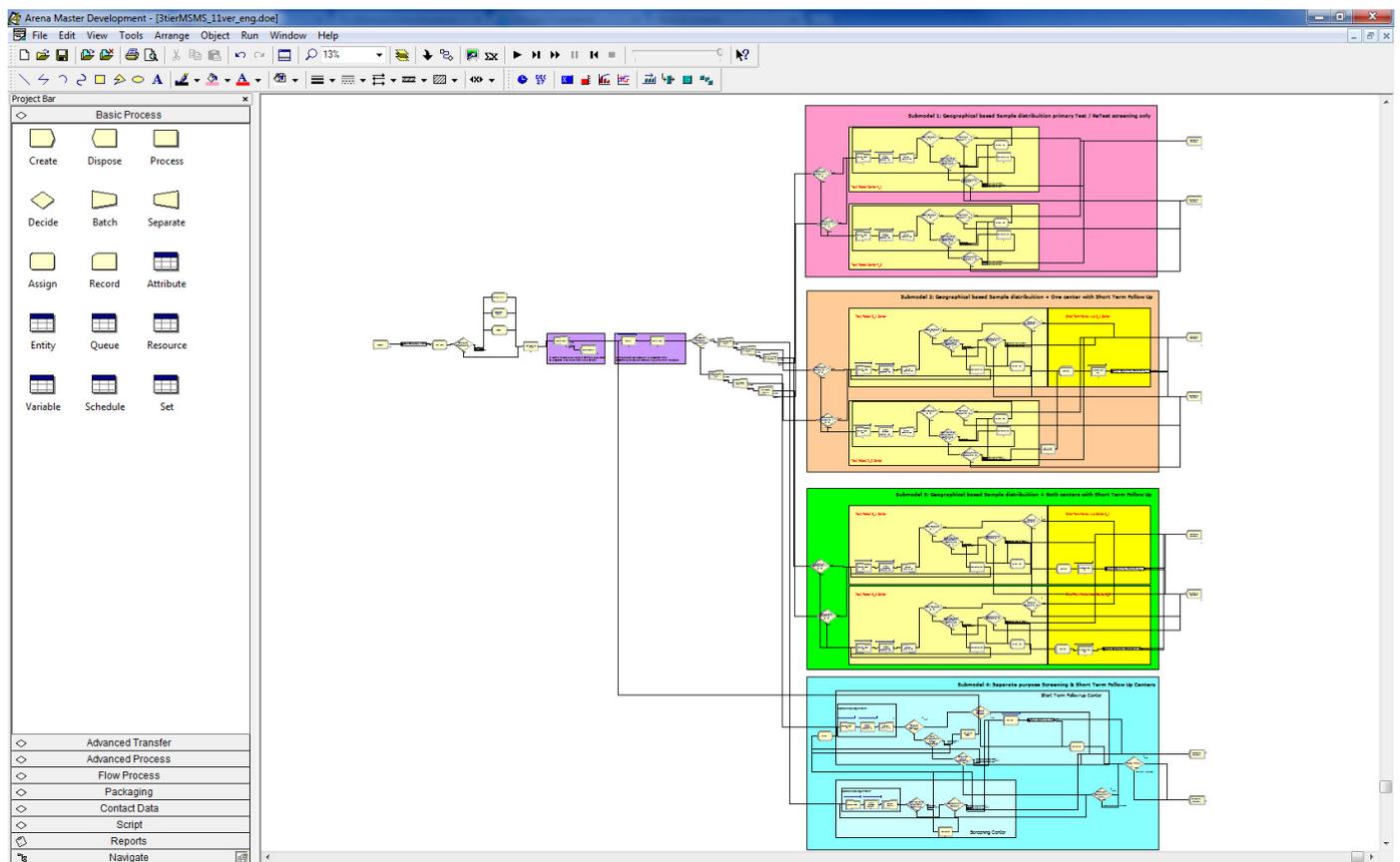


Figure 8. Complete Arena model

The influence of changing environment for path dependence in hierarchical organizations

Arne Petermann
Berlin University of Professional Studies
Department of Business and Management
Katharinenstraße 17-18, 10711 Berlin, Germany
E-mail: arne.petermann@fu-berlin.de

Alexander Simon
Berlin University of Professional Studies
Department of Business and Management
Katharinenstraße 17-18, 10711 Berlin, Germany
E-mail: simon@saw-instruments.com

KEYWORDS

agent based modelling, social simulation, path dependence, path breaking, organizational studies, business strategy, technology strategy

ABSTRACT

The following paper aims at describing how path dependent hierarchical organizations are affected by a changing environment. The results of the latest research in this field (Petermann et al. 2012) analyzed path dependency of norms and institutions in different kinds of hierarchical organizations and the impact of leadership within this process. The results were produced for stable environments only. Agent based simulation was applied as research method. In order to examine how this process evolves when the organizational environment in changing the model M1 (Petermann et al. 2012) will be enhanced. The goal is to simulate the impact of external influences to the norms emerging within the organization. The research described is not yet completed so this paper characterizes first thoughts and possible simulation approaches to tackle the problem at hand.

INTRODUCTION

Nowadays most organizations have to deal with a changing environment. From the organizational point of view a changing environment can be seen as disturbances from outside that forces the organization to adapt very fast. If the organization fails to do so, it may fall back or even be eliminated from the competition. This is especially risky when new technologies flood the market and companies have to react. Examples may be found by taking a closer look at companies like Loewe or Nokia. Loewe missed the technology change on the TV market from the CRT displays to the new LCD-based flat screen technologies. In fact, Loewe builds first rate CRT displays even today, but the market is no longer asking for TVs of this category anymore. Thus Loewe appears ignorant of market realities. The high technical level of their obsolete skills is disguising the internal view of the environment, in this case innovations on the TV market. In the end they were bought by some investors just in time, but their previous ignorance almost led them into bankruptcy.

Comparable to Loewe is Nokia and their behavior on the smartphone market. Nokia was one of the pioneer companies on the mobile phone market, but they did not react adequately to new mobile trends. Just like Loewe, Nokia suffered immensely when other suppliers like Apple (trendsetter of the smartphone trend) and Samsung captured the market. By now the mobile phone division of Nokia has been bought by Microsoft and they slowly get back on track. The questions that arise are: why do companies sometimes need to get hit so hard from external influences until they see that they have to change? How fierce do these influences need to be?

In the following research the model M1 (Petermann et al. 2012) that for reasons of simplicity was built on the assumption of a stable environment will be extended with a new variable, that will include environmental change into the model.

LITERATURE REVIEW AND RESEARCH QUESTION

The theoretical concept for the behavioral analysis described above is called theory of path dependence. The concept of that theory was first described by David (1985). He dugged into the history of the "QWERTY"-keyboards from their first steps in the 19th century until 1985. This alignment of characters has been dominant till today for nearly 100 years. In the early 1930s the alternative "DVORAK" keyboard layout was developed. This at the time new technology was clearly a better and more efficient solution for keyboard layout than the incumbent. These keyboards, however, were not able to become a serious competitor to "QWERTY"-keyboards. David wanted to know why this was so, i.e. what kinds of effects were responsible for the domination of the established keyboards.

Based on David's findings, Arthur (1989) described this market behavior with a model of two technologies (A and B) coming to the market at the same time and fighting for the adoption of the customers, called agents. At the beginning both technologies have the same possibility to get adopted. For the first time in the history of the path dependence debate, Arthur coined the definition of the historical small events increasing returns and contingency. These events are responsible for the start of the path process and lead to a lock-in of

the technologies A or B. Figure 1 (Arthur 1989: 120) illustrates this behavior. When B is locked in, A is completely eliminated from the market.

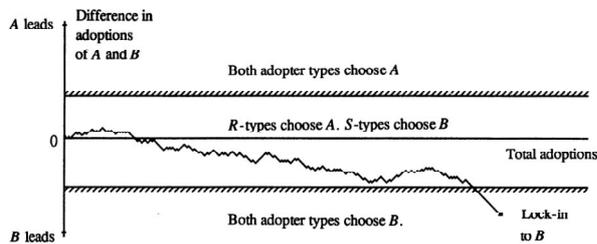


Figure 1: Increasing returns adoption: a random walk with absorbing barriers

Path dependence in the organizational context

To transfer the theory of path dependence to an organizational context a different view of Arthur's description is needed. In organizations and social systems history always matters, and due to the ongoing variations in behavior, the lock-in on markets has peculiar characteristics. There is less adoption behavior, hence development phases deviate from purely technological path dependence. To capture organizational path dependence, Sydow et al. (2009) developed a model which describes this advanced concept of organizational path dependence. In this model the path development is split into three different phases. Figure 2 (Sydow et al., 2009: 692) shows the concept of this model.

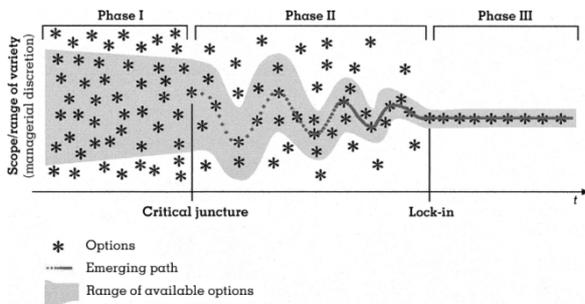


Figure 2: The Constitution of an Organizational Path

Phase 1: Preformation Phase

In this phase the decision the participants are able to make are relatively open. Influences at this time could be historical events, "history matters", routines, and the existing culture of the organization. In the beginning the participants already have an idea of thinking and behaving in their daily environments. Koch (2007: 286) described imprinting circumstances of an organizational culture in this context. Therefore the decisions that will be made in the future are already not completely open. In figure 2 all options are symbolized by the black stars,

but only the stars in the grey zone are available options for the organization.

Phase 2: Formation Phase

In this phase the path begins to emerge. The step from phase 1 to phase 2 is called "critical juncture". At this point an unknown or virtually unrecognizable event from the past leads to the organizational path formation (Sydow et al., 2009: 693). These events are described as "small events". Furthermore self-reinforcing effects triggered by these small events narrow the path. Not every small event is able to trigger such a process, but some will do so.

Phase 3: Lock-In Phase

The reinforcing effects have now taken the lead and reduce the scope of choices. In this phase the organizational path has become locked in. The lock-in state may, in an unfortunate case, be an inefficient one which disables the organization's ability to change and adopt more efficient solutions to the problems at hand. As described above, Loewe appeared locked-in to such an unfortunate state. At first the state was very efficient, but when the market changed Loewe's technology was not needed anymore and thus the state lost its efficiency causing severe problems for the organization.

RESEARCH QUESTIONS

We are interested in the impact of a changing environment to organization that undergo path dependent processes. In historical analysis scholars have shown many examples of organizations that were able to adapt in the light of changing environment while other are stuck in a lock-in state, unable to change even when the necessity to change was obvious. Our model aims at describing hierarchical organizations that undergo a path dependent development in a changing environment. Will they be able to adapt or do they stick to the path? What can we learn about this process applying simulation methods? How should an organization be structured to be able to adapt in the light of dramatic changes in the environment?

METHOD

In modern social and management sciences the method of simulation modeling has been accepted since the early 1990s, like Harrison (2007: 1232) figured out. When complexity and non-linearity of social systems makes it hard or impossible to develop mathematical solutions, simulation models are a good choice to describe the whole system and its development (Gilbert et al 2005: 16). 'Simulation is particularly useful when the theoretical focus is longitudinal, nonlinear, or processual, or when empirical data are challenging to obtain' (Davis et al, 2007: 481). On the other hand it is important to know that the method of simulation cannot replace empirical or analytic methods, but it can provide

first insights for other social research methods to be applied later on.

The basic model

The basic of this research is the simulation model M1 Petermann et al. (2012) developed in his simulation study about the competing powers of self-reinforcing dynamics and hierarchy in organizations. The theory of that model is the simulation of a norm A and a norm B in an organizational hierarchy structure and to answer the question which norm will be adopted by most of its members. Every member of the organization is represented as an agent. These agents are able to decide whether to adopt norm A or norm B.

Agents decision algorithm to adopt A or B

To implement this technically, the agents need to be forced to adopt a norm. Therefore the force-to-act variable FTA is defined (Petermann et al. 2012: 726).

$$FTA_j = E_j * V_j = E_j * \left[\sum_{k=1}^m (V_k * I_{j,k}) \right] \quad (1)$$

V_j describes the connection of individual and organizational goals according to Vrooms (1964) expectancy theory. $E_j \in [0,1]$ represents the subjective probability of each agent's decision. This variable represents the "small events" of the organizational path dependence theory. To implement this in the algorithm, the strictly monotonously increasing function

$$f_{M,c(x)} = e^{m*c*x*1,5} + i(y) * li \quad (2)$$

is used in the simulation to determine V according to equation (1) with $M \in \{A, B\}$, $m = 1$ for $f_{A,c(x)}$ and $m = -1$ for $f_{B,c(x)}$. The variable c represents the reinforcing effects and is generated by the actual spread $\epsilon \in [-1, 1]$ which is a variable that characterizes the state of the system that is either dominated by agents who all choose A (spread = -1) or agents who all choose B (spread = 1) or at some state between these extreme case (spread between -1 and 1). The factor $i(y)$ sets the value of li in the correction path direction. This could be 1 or -1. At the beginning of the simulation the spread is 0 (meaning there are equally large groups of agents choosing A and B in the beginning of the simulation). The lock-in state is nearly 1 for A or nearly -1 for B after a defined amount of time (measured in ticks). The misfit costs are described by x . The leadership impact variable li , which makes the simulation of a hierarchy organizational structure possible, is affecting every agent according to what norm this agent's supervisor prefers.

Under these conditions the agents choose an adoption for A, when

$$FTA_A(x) = E_1 * f_{A,c}(x) > FTA_B(x) = E_2 * f_{B,c}(x) \quad (3)$$

and otherwise B if $FTA_B > FTA_A$.

Simulation of an external impact

Now an external impact has to be implemented in the FTA function to see whether or not this will have an effect of breaking the organizational path. Therefore, equation (2) needs to be extended with an additional value.

$$f_{M,c(x)} = e^{m*c*x*1,5} + i(y) * li + s(z) * ei \quad (4)$$

The variable ei represents the external impact from the changing environment. The factor $s(z)$ is only used to set the correct direction, which depends on the actual path. The value generation of that variable needs to be clarified in the next step. While all variables in the equation are generated by the simulated organization itself, ei is triggered from an external mechanism. When there is no external impact, ei is equal to 0 and behaves neutrally. The question of how the model reacts after the lock-in has occurred is highly interesting. Are there any options to "reset" the norm distribution of the organization? The goal here is to find out about the behavior of the organization regarding the external impact. Are its intensity, its continuity, or a mix of both able to break the path? Every agent in the system is subject to the same external impacts. We assume that environmental influences have the same strength throughout the organization.

Expected Results

With the variable of the external impact, a path breaking effect should be realized. Figure 3 shows a possible behavior during the path formation phase as inferred from first simulation attempts. The path was prompted by an external impact to "reset" the building process.

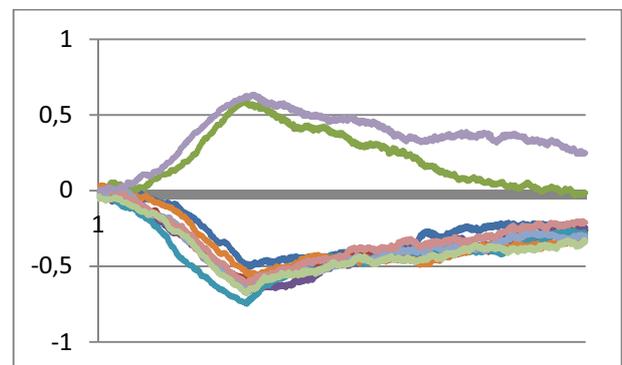


Figure 3: Impacts during path formation phase

The next interesting analysis explains the behavior of an organization with an external impact, when the path has

become locked in. Figure 4 shows some possible path directions similar to figure 3 from first simulation attempts

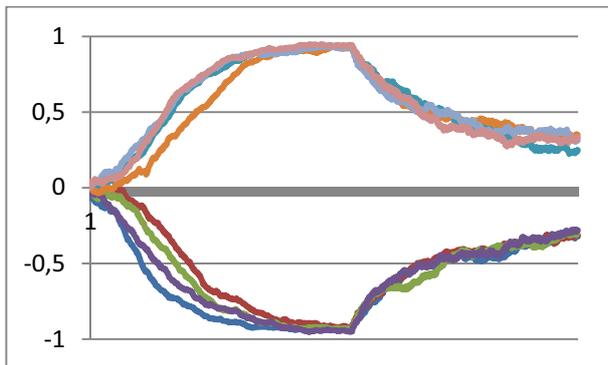


Figure 4: Impact after lock-in

Another interesting part could be the behavior of an organization in cases where the external impact is present right from the beginning on, so before the system eventually becomes locked-in. The question here is, whether the organization is able to properly focus on a norm under such conditions at all.

REFERENCES

- Arthur, B. 1989. "Competing technologies, increasing returns and lock-in by historical events." *The Economic Journal*, 99 (March 1989), 116-131.
- David, P. 1985. "Clio and the Economics of QWERTY." *Economic History*, Vol. 75, No. 2, 332-337
- Davis, P, Eisenhardt, K. and Bingham, C. 2007. "Developing Theory through Simulation Methods." *Academy of Management Review*, Vol. 32, No 4, 480-499.
- Gilbert, N. and Troitzsch, K 2005. "Developing Theory through Simulation Methods." 2. ed, *Berkshire: Open University Press*
- Harrison, J, Lin, Z. and Carroll, G. 2007. "Simulation Modeling In Organizational And Management Research." *Academy of Management Review*, Vol. 32, No. 4, S. 1229-1245.
- Koch, J 2007: *Strategie und Handlungsspielraum: Das Konzept der strategischen Pfade. Zeitschrift Führung + Organisation*, 76(5): 283-291
- Sydow, J, Schreyögg, G. and Koch, J. 2009. Organizational path dependence: Opening the black box. *Academy of Management Review*, Vol. 34, No. 4, 689-709
- Petermann, A., Klaußner, S., Senf, N. 2012: Organizational Path Dependence: The Prevalence of Positive Feedback Economics in Hierarchical Organizations, in: Troitzsch, K. G., Möhring, M., Lotzmann, U. (Hrsg.): *Proceedings 26th European Conference on Modeling and Simulation*, Koblenz 2012, 721-730.
- Vroom, V.H. 1964. *Work and motivation*, New York

Dr. Arne Petermann was MBA Program Director at the Berlin University for Professional Studies from 2011 until 2013. As a visiting scholar, he is currently teaching organization science and scientific simulation methods in the PhD-program at the School for Business



and Economics at Freie Universität Berlin. He is founding President and CEO of Linara GmbH, an international HR agency specialized in the transnational European healthcare sector. His research focuses on organization science, path-dependence theory, entrepreneurship, and social simulation, especially agent-based modeling. His research results have been presented at the leading international science conferences in his field, including the Academy of Management and the American Marketing Association Educators Conference where, moreover, his work was recognized with a best paper award. His work is published in books and peer-reviewed journals.

Alexander Simon was born in Rheinfelden, Germany and studied electrical engineering at the University of Applied Sciences in St. Augustin, Germany. He graduated in 2011. After first professional



experience in the area of software development with focus on the .NET framework in telecommunication and biotech industry, he works in the tax advisory division of Ernst & Young since 2014. During his MBA program at the Berlin University for Professional Studies he became acquainted with path-dependence theory.

HCCM - A CONTROL WORLD VIEW FOR HEALTH CARE DISCRETE EVENT SIMULATION

Nikolaus Furian
Michael O'Sullivan
Cameron Walker
Department of Engineering Science
University of Auckland
Auckland, New Zealand
E-mail: nfur003@aucklanduni.ac.nz

Siegfried Vössner
Institute of Engineering and Business
Informatics
University of Technology Graz
8010 Graz, Austria
E-mail: voessner@tugraz.at

KEYWORDS

Health Care, Discrete Event Simulation, Activity Scanning, Simulation Control

ABSTRACT

The classical world-views of discrete event simulation (DES) are event scheduling, activity scanning and process interaction. A fourth approach, the three-phase method, extends activity scanning and is often regarded as another world-view. These world-views provide the theoretical framework for applying DES in practice. However, in health care simulation, practitioners often face modeling challenges where the concepts and methodologies described by these world-views are not able to reflect either the dynamics or the entity flow of the system being modelled. This leads to individualized approaches and solutions that do not build on a unified and standardized theoretical basis. In this paper we present an extension to the activity scanning world-view based on the needs of the health care sector that uses hierarchical control structures as a more general, flexible and powerful tool to define health care DES models. To demonstrate the strength and potential of the approach two ongoing simulation studies are briefly outlined and the benefits of using the new world-view for these models is discussed.

INTRODUCTION AND MOTIVATION

In DES the underlying system is changed at discrete points in time. How these changes are modeled, controlled and triggered is dependent on the sub-domain of discrete event simulation, also known as the *world-view*, being utilized. The classical world-views are: event scheduling; activity scanning; and process interaction. However, the three phase method, which is an extension to the activity scanning approach, is also often referred to as a world-view. Much work has been done to formalize, evaluate, compare and transform the different world-views, as in (Balci 1988; Birta and Arbez 2007), or more recently (Overstreet and Nance 2004).

Recently the need for a uniform description of discrete event models has been identified (Robinson 2006; Balci

et al 2008; Robinson et al. 2010). However, a uniform descriptive language, that is generally accepted and used, requires that the underlying simulation theory can be used to model the majority of simulation projects. Unfortunately, although most published models in the health care sector, are based on one of the classical world-views, these models often include customized features that are not included in that world-view, to fit the requirements of the specific system being modeled, as for example (Günel and Pidd 2011) who generate multiple “mini-doctors” out of a doctor to imitate multi-tasking.

In this paper we will outline some of the major drawbacks of the activity scanning and three phase method world-views for health care simulation followed by the description of concepts that attempt to resolve these shortcomings and lead to more generality and flexibility within activity based DES. The paper is structured as follows. In the following section we discuss the three phase method world-view and the ABCmod framework that developed from it. In the next section we revisit the concept of activities, followed by a definition of the hierarchical control mechanism. Section *Time Advancement* briefly describes time advancement for the proposed method. The paper concludes with two case studies from health care, suggestions for further research and final remarks.

THE THREE PHASE METHOD AND ABCmod

In the original three phase method, see for example (Pidd 1995), the only concept of control is that of conditional activities in the entity flow within a model. Based on this method, (Birta and Arbez 2007) introduced the most structured and deliberate modeling theory and language for DES, the ABCmod framework. Besides conditional activities, they identify queues as core elements of any activity based DES to control the entity flow.

Although, queues and conditional activities are able to reflect less complex, more rigid systems, such as those that often occur in manufacturing problems, they often fail to represent health care systems with complicated, dynamic dispatch policies, resource flows and entity relationships. The need for more flexible control

structures in DES has been addressed previously by (Pratt et al. 1991), who outlined the importance of separating control and informational elements of a model. For health care modeling, (Hay et al. 2006) pointed out that sophisticated dispatch systems and the skill set of staff members drive the process rather than patients in simple queues. A more advanced approach has been proposed by (Lim et al. 2013), where entities can assign tasks to other entities. Definitions of structured control mechanisms for process and job oriented simulation has been introduced by (Raunak et al. 2009), (Mes and Bruens 2012) and (Robinson et al. 2006).

The standard approach based on conditional activities and queues can be briefly summarized: as soon as a condition evaluates to true an entity is chosen from a queue (e.g. first-in-first-out or priority based) and an activity is triggered. This approach requires the independence of queues from each other and a relatively rigid mapping of resources to queues. However, health care scenarios often consist of: more complicated dispatch methods; resources that regularly change their role to consumers and vice versa; time-dependent priorities; and skill-level dependent job allocations. Furthermore, one often faces a high degree of interruptions and concurrent tasks. Therefore, dispatch policies have to select from a pool of future and uncompleted jobs, taking into account a very heterogeneous set of factors including: time related priorities; task related priorities; degree of completion; future assigned tasks; skill levels; and individual preferences.

When modelling using classical queuing with conditional activities, practitioners quickly encounter the limitations of this approach and create customized workarounds. The simulation control concept presented in this paper replaces conditional activities and queues by a hierarchical control tree. This tree consists of nodes that represent control units with activities at the leaves. In this approach more general rule sets replace dispatch via queues and conditions, thus providing greater flexibility when modelling.

Furthermore, the hierarchical control world-view enables the integration of optimization within simulation. The classical approach to the integration of simulation and optimization uses a simulation model to evaluate an objective function of an optimization model. Thereby, the simulation model and its implementation act as a black box. Hence, both systems, the simulation and optimization, are encapsulated and only interact via the exchange of parameters and objective values. In addition to this standard approach, the hierarchical control world-view, presented in this paper, allows optimization techniques to be elegantly embedded within a simulation model to control the simulation flow itself. Although, optimization has been embedded within

simulations in previous applied research, a uniform theoretical basis for integrating optimization techniques to control entity flow is still missing in literature.

REVISITING ACTIVITIES

Entities and their behavioral artifacts are two of the main elements of any DES. Behavioral artifacts represent the flow of entities and their relationships and interactions with each other. Usually, two types of behaviors are defined, events (or actions) and activities. *Events* cause instant changes to the simulation model's state, for example the arrival of an entity. The classical definition of *activities* is that they stretch over a certain period of time, consist of at least two events: the start; and end event; and represent a purposeful task.

In the classical world-view literature, events and activities are classified as scheduled, conditional or sequential activities depending on their trigger mechanism.

Scheduled events and activities occur at designated simulation times, independent of the simulation model's state. Further, they represent the mechanism to advance time during simulation. Activity scanning and the more advanced three phase method both scan through all *conditional* events and activities each time the simulation clock progresses and triggers an event or activity if a condition evaluates to true. *Sequential* events and activities are triggered immediately after the termination of a preceding behavioral artifact. In particular, activities are denoted as scheduled, sequential or conditional if their start events are of the corresponding type.

In the hierarchical control world-view scheduled and sequential behavior is handled in the standard way, but a single class for conditional behavior is not considered sufficient to enable general, flexible simulation models, particularly with respect to simulation control and advanced dispatch.

To demonstrate the shortcomings of the standard classification of activities and to illustrate the proposed hierarchical control world-view we will use a simple patient transport simulation. Patients located at a ward require escorting to a diagnostic facility. We assume that these requests occur randomly according to some kind of stochastic process. They stay in the ward until they are guided to an empty diagnostic site, in the diagnostic facility, by an orderly (we assume that there is only one orderly in the model), an activity called escort-to. Note that there are only a limited number of diagnostic sites available in the facility. In the diagnostic site patients receive some sort of diagnostic assessment and are subsequently escorted back (called escort-back) once the orderly is available, i.e., escorted by the orderly from the diagnostic site to the ward. After being escorted back,

the patient stays in the ward until she leaves or needs another service. Furthermore, the orderly moves by herself from the diagnostic facility to the ward if no patients are ready for being escorted back, there are patients waiting in the ward for transport and at least one diagnostic site is available. On the other hand, the orderly moves by itself from the ward to the diagnostic facility if no patients are waiting for transport and at least one patient is currently diagnosed or has finished diagnostic assessment.

According to the standard classification of activities, escort-to, escort-back and the orderly moving by herself whether from the ward to the diagnostic facility or vice versa are conditional activities. However, there is a significant difference in the trigger mechanisms of these activities.

First, both escorting (to and back from the diagnostic facility) activities are *requested* by the patient's behavior. Although, the activities are handled and triggered by the control tree, they are motivated by the patient's behavioral path. Hence, we refer to them as *requested* activities.

On the other hand, the orderly moving by herself has different motivations and trigger mechanisms. The behavioral path of the orderly does not trigger an empty move. These moves are determined by the control policies of the model and triggered by the hierarchical control tree, hence are called *controlled* activities.

However, there can be a significant difference in the trigger mechanisms of controlled activities. When a patient needs a diagnostic assessment she requests being escorted to the diagnostic facility. When the hierarchical control tree assigns the orderly to this request, it will trigger the orderly to move by herself to the ward if necessary. Similarly, the control tree will trigger an orderly to move to the diagnostic facility when it assigns the orderly to a request for escorting a patient back after the patient's assessment is finished. However, if no patient has filed an escort-request and at least one patient is being assessed, but has not completed this assessment (in the diagnostic facility) then there is no request to assign the orderly to and, hence, trigger the orderly to move by itself. This movement is instead triggered by the hierarchical control tree because it should reduce the waiting time of the patient.

Hence a controlled activity can be triggered:

- 1) in response to a dispatched activity request; or
- 2) to improve the system's performance, e.g.,
 - a. by anticipating future requests; or
 - b. by considering where undispached requests are located.

We distinguish between requested behavior (i.e., requested activities and events) and controlled behavior

(i.e., activities and events that are triggered by the hierarchical control tree), but make the modeling distinction between controlled activities that occur in response to a activity request and controlled activities that occur as part of the control tree's overall plan, see figure 1.

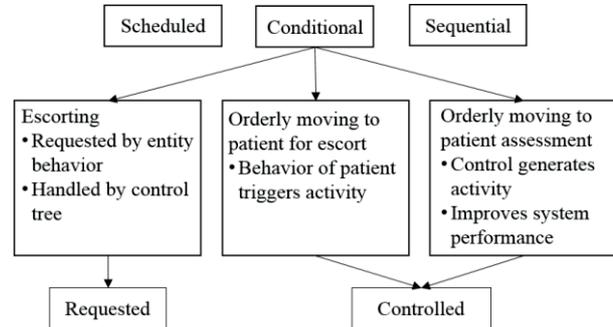


Figure 1: Extended Activity Classification

Note that requests for an activity to take place are themselves elements of the simulation model. As a consequence they may have attributes, for example the time when the request was generated, who generated it and/or an associated priority.

The definition of providing and consuming roles is part of many concepts for DES, for example in the framework of (Birta and Arbez 2007). Although their definition fits the manufacturing environment reasonably well, it is too rigid and not applicable to health care DES. In the hierarchical control world-view, any roles for entities concerning their state in terms of consuming or providing become irrelevant. For example, a less experienced surgeon assisting in a complicated surgery, as part of her educational program, cannot be unambiguously defined as either a provider (resource) or consumer (entity). On one hand, she assist during the activity and therefore actively contributes to the surgery (i.e., she is a provider). On the other hand, she benefits from the teacher-student relationship and is gaining knowledge from the more experienced surgeon leading the operation (i.e., she is a consumer). In the hierarchical control world-view, the pooling of activity requests combines with the hierarchical control tree to replace entity queues and renders the distinction between providers and consumers unnecessary.

HIERARCHICAL SIMULATION CONTROL

In the previous section one main element of the new hierarchical control world-view was introduced. Conditional activities were re-classified depending on how they were triggered in the system. In this section the other main element of the hierarchical control world-view is defined. The hierarchical control tree is introduced and request handling is described. Systems

modeled using the hierarchical control world-view are hereafter referred to as Hierarchical Control Conceptual Models (HCCMs).

General Concept

One could describe discrete event models in a very informal and colloquial way as “a model that consists of a set of entities and set of tasks that have to be completed”. Furthermore, behaviors and rules define the sequence of activities and the assignment of entities to activities respectively. While manufacturing systems tend to have more rigid structures, health care models often fit the informal description quite well. Patients and staff members request activities of various forms including: treatments; assessments; teaching; organizational work; rounds; meetings; and many more. These requests are hard to assign to queues as many activities require multiple resources, and resources perform multiple activities.

The hierarchical control world-view proposed in this paper follows a different approach. Instead of separating activities into different types and handling them in corresponding queues, activity requests that are associated with the same organizational area, or unit, are pooled together in Requested Activities and Events Lists (RAELs). Additionally, for each organization area, a control unit is defined to handle the corresponding activity requests. These organizational areas can be seen as sub-models of the entire system. The use of a model/sub-model structure has already been identified as a substantial requirement for model reuse and the design of integrated models, see (Zeigler 1987). We propose linking sub-models together into a hierarchical tree structure with control units at all nodes except for the leaves which are requested or controlled activities, as shown in figure 2. It is important that the control tree and requested activities are designed in an unambiguous way. In particular, the same requested activity cannot be the child of more than one control unit. Thus, activity requests are non-ambiguously handled by unique control units.

Control units manage entities, manage requested and controlled activities and handle activity requests. Further, they consist of a set of rules that determine the conditional behavior of the model, and a set of delegates that represent the communication between control units.

The depth of the designed tree is obviously a modeling choice, which has to be made under great care. Too much granularity leads to unnecessarily complex models that are cumbersome to deal with. Whereas too few control structures also can lead to rising complexity in conditions for dispatch causing interactions to get more difficult to model.

From a more technical point of view control units should also provide interfaces to obtain controlled elements,

including entities, current activities and requests. Due to space limitations, a detailed, technical definition of all elements is beyond the scope of this paper and is left for future work.

In the following subsections the two main elements of a control unit's definition, rules and delegates, are explained in more detail.

Rules

Rules are one core element of control units. They replace the conditions of activities in a more structured and centralized form, and can be divided in five different categories:

- **Assessment:** What can be dispatched and/or done?
- **Dispatching:** What should be dispatched and/or done?
- **Control:** What should be done given dispatching decisions and/or the current state of the system?
- **Replacement:** What does not need to be done any more?
- **Custom Rules**

Assessment rules do not differ from common conditions in a centralized form. They only assess which activity requests from the RAEL could be dispatched and/or performed given the current state of the system, particularly in terms of available entities, and put these requests in a list of *possible* requests, referred to as RAEL (i.e., the RAEL holds only requests that can possibly be triggered in the current state of the model).

Dispatch rules decide which of the requests in the RAEL to dispatch and/or do next. In many cases they represent simple queues, but can also model various complex dispatch policies and resource allocation scenarios.

Control rules trigger behavior that is either the result of dispatch decisions (e.g., the orderly moving to perform escorting) or that are designed to improve system performance (e.g., the orderly moving in anticipation of a patient completing diagnostic assessment).

Activity replacement rules determine under which circumstances requests are removed from the RAEL. This could happen for various reasons. For example, if waiting times exceed certain limits entities might leave the model (aka *balking*), or, when an activity has been deferred (e.g., a low priority meeting), the request may have expired (e.g., the time scheduled for the meeting has passed).

Custom rules can be added to model any control functionality not sufficiently represented by the other rule categories.

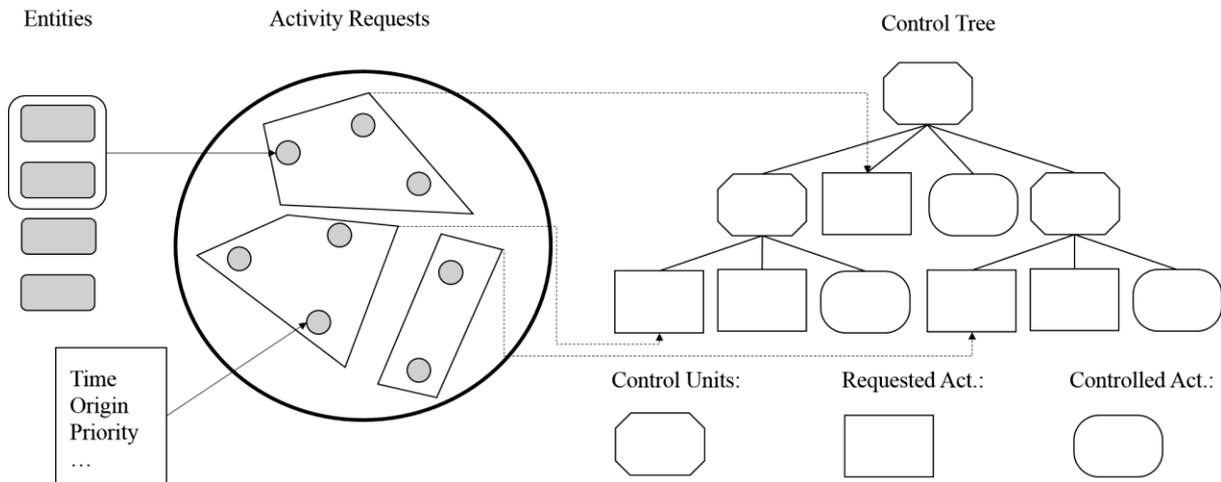


Figure 2: Concept of Hierarchical Control World View

It is important that rules are executed in the correct order. First, rules are considered in a top to bottom approach, meaning that rules of parent control units are checked first and can also differ from rules of child control units. Second, within one control unit the rule sets have to be applied in their particular order: assessment; dispatch; control; and replacement. This will be discussed in more detail in section *Time Advancement*.

Delegates

While rule sets define how control units operate and manage their entities, delegates represent the communication between control units. This messaging should strictly follow the tree structure of the control view. Thereby, bottom-up and top-down messaging is possible. Typical scenarios that require the formulation of delegates could be the report of special circumstances, for example, the occurrence of an emergency; overworked entities; the request for additional entities; or the moving of activity requests up and down the tree structure.

Basically, any kind of custom delegates can be defined. However, it has to be ensured that the receiving control unit is equipped with a corresponding rule set (receiving delegates) to handle it properly.

TIME ADVANCEMENT

In this section an algorithm for time advancement and rule execution is presented. It is based on the classical three phase method, which itself is an extension of the activity scanning world-view. Analogous to the standard three phase method, time advancement is driven by scheduled behavior. Scheduled events, either standalone or the activity start, are held in a list sorted with respect to occurrence time, referred to as the Scheduled Event List (SEL).

Each time a scheduled event is scheduled to start, the simulation clock advances and the different rule sets are used to trigger controlled behaviors. This is done in a top-down manner according to the hierarchical control tree. Thereby, the rule sets are considered in their particular order: assessment; dispatch; control; and replacement. If a requested or controlled event (standalone or activity start) has been triggered the algorithm steps back to the assessment rule set. If no action could be launched the algorithm advances to delegates. Delegates are sent both ways, bottom-up and top-down. Due to the hierarchical tree structures first delegates are sent bottom-up followed by the top-down messaging.

This process is repeated until no behavior was triggered and no delegate was sent or received. Only then the next scheduled event is considered, assuming no stopping criteria has been met. Figure 3 shows the basic principle of the time management and activity control algorithm.

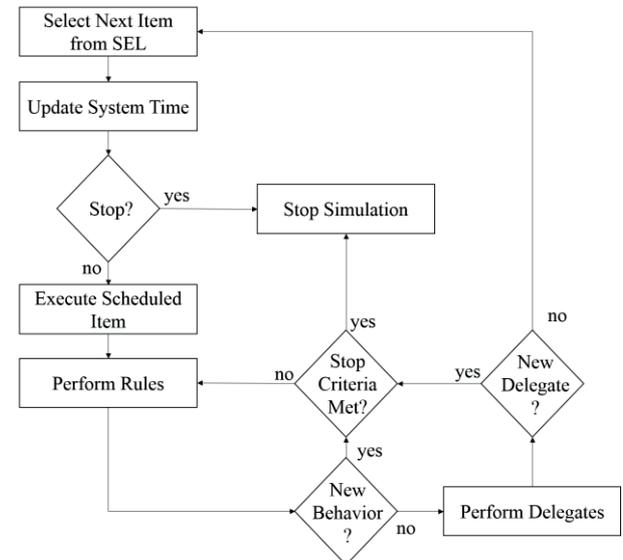


Figure 3: Time Management Algorithm

CASE STUDIES

To highlight the strengths of the proposed method two case studies are briefly discussed. For both detailed HCCM models have been created, including entity definition, formulation of event routines and all rule sets of control units. However, due to the limited space only the main characteristics of the case studies' models are outlined and the benefits of the hierarchical control world-view for these models are discussed.

Patient Transport Simulation

In an on-going simulation project at the University of Auckland, the transit of patients at Auckland City Hospital (Auckland, New Zealand) has been modeled. Patients are transported between wards and treatment facilities. These transits are performed by orderlies and, in some cases, additional nurses if required by the patient's condition. Each patient has a record of scheduled treatments he must undergo and additional emergency transports. The objective of the model is to assign orderlies (and nurses if required) to the requested transits between wards/treatment facilities in an efficient way. The request for transit is launched 40 minutes prior to the actual treatment appointment.

Using the standard queuing based discrete event simulation approach, patients that require a transport would be held in queues, either one global queue hosting all patients, or separate queues for each location in the model. As soon as resources (orderlies and possibly nurses) become available patients would be picked from the queues according to a selection routine.

However, if no other transit queued, after a completed transport orderlies return to their home base and nurses return to their home ward to wait for further assignments to transits. Hence, upon the occurrence of a transit request, usually the required resources are not present at the patient's location. This requires empty moves, i.e. controlled behavior, of both orderlies and nurses. Although the empty moves could be included in the actual transport activity, two major limitations would remain. First, the pre-emption of an empty movement with respect to the occurrence of a higher priority task, e.g. emergency transport, would require the re-insertion of the patient in the request queue at the same position the previously occupied. Second, and more important, selection of patients in queues (waiting for transit) is only triggered upon the availability of resources. However, more sophisticated dispatching routines may also take into account orderlies and nurses that are currently performing a transport. This certainly stretches the boundaries of queuing-based discrete event simulation.

The different rule sets of HCCM models presented in this paper overcome these difficulties. During assessment, jobs that can be dispatched are identified. Dispatching rules assign tasks to orderlies,

independently of their availability and the control rule set triggers empty movements towards a patients and initiates actual transports. Replacement and custom rules were not necessary for this particular model.

The main focus of the project is to investigate different dispatch policies in order to identify an optimal dispatch method. The target of any dispatch routine for orderlies is that patients arrive between 15-0 minutes before the scheduled treatment, with penalties for late arrivals.

The first dispatch routine reflects the current real world strategy and assigns the nearest available orderly (and nurse, if necessary) for a new job. The second policy attempts to estimate which orderly could be at the patient's position earliest, including the time required for finishing other jobs first. The third dispatch routine really demonstrates one of the strengths of HCCM models: the dispatch rule set is "replaced", or "represented", by a Mixed Integer Linear Program (MILP) created from a snapshot of the system's state. Each snapshot leads to an undirected graph, where nodes represent jobs and edges represent the completion of a job plus the empty move to the location of the next job. Based on that graph a multi-travelling salesmen problem with time windows is formulated as an MILP. A solver is then used to solve the problem and the optimal solution dictates the next dispatch decision(s). Obviously optimization methods, such as MILP solvers or heuristics, have been used before to control simulation models. But the hierarchical control world-view provides a theoretical foundation to do so, which, to the knowledge of the authors, has not existed so far.

Cytology Lab Simulation

The Cytology lab study is currently under progress at the University of Auckland in cooperation with staff at LabPLUS laboratory in Auckland, New Zealand. In general, LabPLUS laboratory performs cytology analysis on different samples sent by surrounding health care facilities. These samples are analyzed by pathologists employed by the lab. In addition to sample analysis, pathologists perform a heterogeneous set of other tasks, with varying priorities, both in the lab and located in other facilities. Most of these tasks are scheduled and their date and location are known in advance. However, pathologists also participate in "call-outs", which could be best described as emergency calls from surrounding hospitals. As it is the nature of emergencies they occur randomly and their duration is not known. Other tasks include: meetings, both in and out of the lab with varying priorities and policies for attendance; preparation for meetings; teaching; clinic visits; and organizational work.

One of the main challenges during the modeling process was the large number of interrupted activities. Pathologists constantly have to stop their current, incomplete activity to perform a higher priority task. As

a consequence, while assigning jobs to pathologists, any dispatch routine has to take into account: all pending tasks; all incomplete tasks of pathologists; future scheduled activities (e.g., meetings or clinical visits); waiting times; remaining time windows for task completion (e.g., preparation time for meetings); rostering and shift data; skill levels of pathologists; blocked pathologists for certain activities (e.g., second opinions on samples); and of course the priorities of the tasks themselves. The highly heterogeneous set of activities and their dependencies was difficult to model in commercial software tools. However, a HCCM model could be elegantly stated, including the model description and the definition of dispatch routines that address all the challenges of the system.

It is not possible to outline all challenges overcome during the modeling process in detail here. However, some features that highlight limitations of the standard queuing approach and emphasize the strengths of HCCM models are illustrated in the following.

First, one feature of the model was the inclusion of training sessions for junior pathologists. Each junior pathologist has to participate in a certain number of one on one training sessions provided by a senior pathologist each month/week, depending on her skill level. Hence, with respect to the standard queuing paradigm the junior pathologists would be resources and entities at the same time. This makes independent evaluation and selection routines of queues impossible. By pooling all activity requests, regardless of being generated by a sample, pathologist or other entity more general selection functions can be formulated.

This leads to the second distinct feature of the HCCM model. As requests for different categories of tasks, with varying participants and necessary resources, are pooled together, more general dispatching functions can be defined. All possible tasks (assessment rules) are weighted according to their type, request time, degree of completion and priority. Dispatching functions then match activities and their weights with appropriate pathologists to decide what happens next. Thereby, future dispatched tasks (e.g. meetings) are also taken into account. The advantage is that jobs of different kinds are handled in one structure, and dependencies of queues that would be cumbersome to model, become irrelevant.

An additional challenge was the modeling of some kinds of meeting. While external meetings are modeled in a straight forward way, internal staff meetings are trickier to handle. Staff meetings are scheduled meetings that require all members of the lab to participate, unless they are currently engaged in, or booked in the near future for, a higher priority activity. However, the meeting only takes place when a minimum number of available pathologists (quorum) is exceeded. Otherwise, the

meeting is postponed until enough pathologists become free or the latest meeting end time has been reached. Further, it is possible that a pathologist leaves the meeting earlier, e.g. due to the occurrence of an emergency, and re-joins it later. Obviously, this is a very complex situation, and very difficult to model with standard queuing and conditional activity based conceptual modeling frameworks, in an elegant way. However, the formulation of requests leads to an elegant formulation of the problem. The meeting entity generates an initial request resulting in the assignment. At the same time requests for joining are filed by non-available pathologists. Future, dispatching decisions with respect to corresponding rules allow pathologists to join once they are free. Replacement rules remove requests after the latest start-time of the meeting.

FUTURE RESEARCH

HCCM provides generality, flexibility and modularization that lead to a variety of research opportunities and application possibilities. One such example is the coupling of optimization techniques with the hierarchical control tree to simulate the potential of optimal behavior. However, the two most important goals for the near future are to define a uniform description language for HCCM and provide a software solution to implement HCCM models without great effort. Both tasks are currently under progress at the University of Auckland in cooperation with commercial simulation software providers.

CONCLUSION

In this paper we have argued that the standard theory on DES simulation has some major shortcomings for health care simulation. As a result commercial software tools and applied simulation studies often use individually customized models to fit the requirements of real world applications. This leads to significant barriers for interchangeability and re-usability of models, as well as a scientific discussion on health care DES methods. The hierarchical control world-view, and hence HCCM models, make one step towards resolving the limitations of activity-based DES. Based on an extended classification for conditional activities, a new simulation control mechanism is introduced that replaces the restricted queuing paradigm. Thereby, a much higher degree of generality and flexibility in modeling is realized. Further, the modular structure of the approach allows the integration of optimization techniques to control the simulation flow. Two applied problems were briefly introduced to demonstrate that very complicated problems, for which the authors have previously struggled to find satisfactory representations using standard approaches and commercial software tools, could be elegantly described as HCCM models.

ACKNOWLEDGEMENTS

The first author thanks the **Austrian Science Fund (FWF): Project Nr. J3376-G11**, for funding his research in New Zealand.

REFERENCES

- Balci, O. (1988). The Implementation of Four Conceptual Frameworks for Simulation Modeling in High-level Languages. *Proceedings of the 20th Conference on Winter Simulation*, pp. 287--295.
- Birta, L. G., & Arbez, G. (2007). *Modelling and Simulation: Exploring Dynamic System Behaviour*. Springer.
- Günel, M. M., & Pidd, M. (2011). *ACM Transactions on Modeling and Computer Simulation: DGHPSIM: Generic Simulation of Hospital Performance*. ACM, (S. 1-22).
- Hay, A. M., Valentin, E. C., & Bijlsma, R. A. (2006). Modeling Emergency Care in Hospitals: A Paradox - The Patient Should not Drive the Process. *Proceedings of the Winter Simulation Conference, 2006. WSC 06* (S. 439-445). Winter Simulation Conference.
- Lim, M. E., Worster, A., Goeree, R., & Tarride, J. E. (2013). Simulating an emergency department: the importance of modeling the interactions between physicians and delegates in a discrete event simulation. *BMC Medical Informatics and Decision Making*, 13(59).
- Mes, M., & Bruens, M. (2012). A generalized simulation model of an integrated emergency post. *Proceedings of the 2012 Winter Simulation Conference (WSC)* (S. 1-11). Winter Simulation Conference.
- Overstreet, C., & Nance, M. C. (2004). Characterizations and Relationships of World Views. *Proceedings of the 36th Conference on Winter Simulation* (S. 279-287). Washington, D.C: Winter Simulation Conference.
- Pidd, M. (May 1995). Object-Orientation, Discrete Simulation and the Three-Phase Approach. *The Journal of the Operational Research Society*, 46(3), S. 362-374.
- Pratt, D. B., Farrington, B. A., Basnet, C. B., Bhuskute, H. C., Kamath, M., & Mize, J. H. (1991). A framework for highly reusable simulation modeling: separating physical, information, and control elements. *Proceedings of the 24th Annual Simulation Symposium, 1991*, (S. 254-261).
- Raunak, M., Osterweil, L., Wise, A., Clark, L., & Henneman, P. (2009). Simulating patient flow through an Emergency Department using process-driven discrete event simulation. *ICSE Workshop on Software Engineering in Health Care, 2009, SEHC '09*, (S. 73-83).
- Robinson, S. (2006). Conceptual Modeling for Simulation: Issues and Research Requirements. *Proceedings of the 38th Conference on Winter Simulation* (S. 792--800). Monterey, California: Winter Simulation Conference.
- Robinson, S., Taylor, S., Brailsford, S., & Garnett, J. (2006). Building Communicative Models – A job oriented approach to manufacturing simulation. *Proceedings of the 2006 OR Society Simulation Workshop*.

Robinson, S., Brooks, R., Kotiadis, K., & Van Der Zee, D.-J. (2010). *Conceptual Modeling for Discrete-Event Simulation*. Boca Raton, FL, USA: CRC Press, Inc.

Zeigler, B. (1987). Hierarchical, modular discrete-event modelling in an object-oriented environment. *SIMULATION*, 49(5), S. 219-230.



MICHAEL O'SULLIVAN is a Senior Lecturer in the Department of Engineering Science at the University of Auckland, New Zealand. He received his PhD in Management Science and Engineering from Stanford University. He works on Operations Research (OR) modeling, applications and computation. His current focus is the use of OR to build intelligent cloud systems and improve health care systems. His e-mail address is: michael.osullivan@auckland.ac.nz



NIKOLAUS FURIAN is a visiting post-doc at the University of Auckland. He obtained his degree in Technical Mathematics at the University of Technology Graz in 2008. During his PhD program he focused on applied operations research topics, especially bin packing. Currently he working on simulation and optimization applications for health care sector. His e-mail address is: nfur003@aucklanduni.ac.nz.



CAMERON WALKER is a Senior Lecturer in the Department of Engineering Science at the University of Auckland, New Zealand. He received his PhD in Mathematics from the University of Auckland. He works on computational analytics, focusing on statistical modelling, simulation and optimization. His e-mail address is: cameron.walker@auckland.ac.nz



SIEGFRIED VÖSSNER holds a PhD degree in Engineering Sciences from Graz University of Technology. Until 1999 he was a postdoctoral fellow and visiting scholar at the Department for Engineering Economic Systems and Operations Research at Stanford University, USA. After being a project manager for McKinsey&Company he became professor and chairman of the department of Engineering- and Business Informatics in 2003 and is currently Vice-Dean of the School of Mechanical Engineering and Economic Sciences of Graz University of Technology. E-mail: voessner@tugraz.at

TOWARDS A SIMULATION MODEL OF PARTNER-SPECIFIC ABSORPTIVE CAPACITY AS A PATH DEPENDENT SELF-REINFORCING MECHANISM IN B2B RELATIONSHIPS

Tobias Grossmann
Freie Universität Berlin
Marketing-Department
Otto-von-Simson-Str. 19, 14195 Berlin, Germany
E-mail: tobias.grossmann@fu-berlin.de

Arne Petermann
Berlin University of Professional Studies
Department of Business and Management
Katharinenstraße 17-18, 10711 Berlin, Germany
E-mail: arne.petermann@fu-berlin.de

KEYWORDS

Path dependence, self-reinforcing mechanism, lock-in, partner-specific absorptive capacity, business-to-business relationships, modeling

ABSTRACT

This study explores the influence of partner-specific absorptive capacity as a self-reinforcing mechanism in the context of the theory of path dependence. For this purpose, theoretical foundations of path dependence and absorptive capacity are reviewed. Following, a path model of partner-specific absorptive capacity for business-to-business relationships is developed. While this paper concentrates on the theoretical foundations of an integrated path model for partner-specific absorptive capacity, the results can be used as a starting point for simulation research in the future. In that sense, this paper shows an early stage simulation research project.

INTRODUCTION

A better understanding of the phenomenon of path-dependent processes is in the interests of academia and real-world practice (Sydow et al. 2009). Efforts to concretize the theory of path dependence have motivated various research approaches since the theory first originated about thirty years ago. In the process, vivid examples have been presented so far, and different attempts have been made to arrive at a general definition.

A promising development is also apparent in the concept of absorptive capacity. The relevant literature on the ability to absorb knowledge repeatedly points to this construct's path-dependent character, but does not examine it in further detail. This reveals an exciting and largely unresearched overlap between the two concepts, an area that requires more in-depth analysis (e.g. Mallach 2012). Although many scholars argue that the concepts of path dependence and absorptive capacity are somehow linked, an integrated model that combines the prominent features of both processes has yet not been developed.

To be able to observe knowledge transfer from one organization to the other, the business-to-business marketing context is an interesting area for our

investigation. Business relationships between buying and selling firms can be generally defined as a non-accidental sequence of market transactions between independent market actors (Kleinaltenkamp and Ehret 2006).

Restricting business relationships between a selling and a buying firm solely to the exchange of market transactions does however seem shortsighted. Often, information sharing plays a central role in market transactions. This includes areas as knowledge of technical details or knowledge about reciprocal net benefits (Brennan et al. 2007). In many cases know-how related to problem solving is also exchanged within the dyadic partnerships (Kleinaltenkamp 1997).

For Van den Bosch et al. (1999) flows of knowledge with regard to products, services, production processes, and market characteristics are important. In addition, Von Hippel (1988) found that knowledge transfer between customers and providers is an outstanding source for innovative ideas. Moreover, Plinke (2000) emphasizes that information sharing enables the supplier to create superior products and/or services for its customers. In the dyadic business relationship, the partners' ability to absorb and share new knowledge is the main driver in learning processes. Furthermore, partners can leverage what they have learned to identify ways to improve the quality, reliability, and speed of knowledge transfer in the future (Chen et al. 2009). Also, inter-organizational learning has a positive effect on the performance of those involved (Gulati and Sytch 2007) while also allowing them to remain competitive (Dyer and Nobeoka 2000).

The ability of business firms to absorb and process external knowledge from its partner is therefore of great importance. We argue that a partner-specific capacity is crucial for these processes. Understanding how mechanisms of absorptive capacity work and how these mechanisms are connected to path dependence is at the heart of this paper. We will build a theoretical model that can be used for formal modeling and simulation research in the future.

The remainder of the paper is structured as follows: In Section 2, we present the theoretical framework by reviewing the literature on path dependence and absorptive capacity. Then we show the connection

between path dependence and absorptive capacity. In Sections 3, we develop a path model of partner-specific absorptive capacity in business-to-business relationships. The last section discusses results and future research directions towards formal modeling and simulation research.

LITERATURE REVIEW

Path Dependence Theory: How History Determines Our Future

The notion of path dependence basically highlights a historical process: initial decisions increasingly restrain present and future choices thereby challenging the a-historical rational choice view. David (1985) initiated the discussion on path dependence from an economic perspective. Within his historical studies he explored the puzzling persistence of the QWERTY keyboard technology and tried to answer the question why an inferior standard was maintained although superior technological innovations were available at different points in time. His exploration surfaced underlying self-reinforcing processes which increasingly rigidified the technological standard (David 1985). Arthur (1989) has formalized the theory of path-dependent processes highlighting the critical role of increasing returns.

In conclusion path dependence has been described as self-reinforcing processes characterized by non-predictability, nonergodicity, inflexibility, and potential inefficiency (Arthur 1989; David 2001; Pierson 2000). More precisely, path dependence is not predictable at the beginning; various alternatives are possible. At a later stage due to self-reinforcing effects the scope of action increasingly narrows and finally leads into a dysfunctional trap, inhibiting the organization to deviate from it. Accordingly, the state of path dependence can be conceptualized as the outcome of a dynamic process that is driven by at least one self-reinforcing mechanism. It proved useful to differentiate this process into three distinct phases for characterizing the sequence of varying regimes: preformation phase, formation phase, and lock-in phase (Sydow et al. 2009). Figure 1 illustrates all three stages.

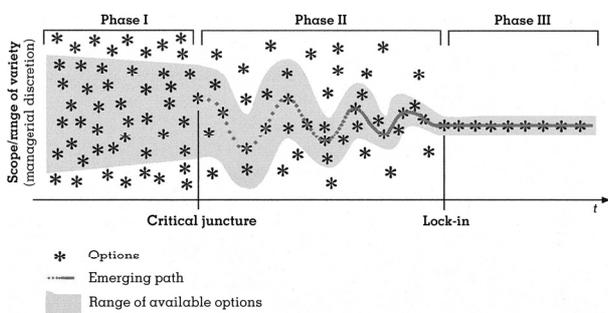


Figure 1: Three Stage Model of Path Dependence According to Sydow et al. (2009)

Over time a path constitutes a restriction of choice for a decision-making system. While choice is not restricted at the beginning, the situation gets more and more restricted in the process, the emerging solution at the critical juncture amounts to a path which increasingly binds subsequent decision making. To put it differently, the emergent solution gets reproduced again and again; when other solutions become unavailable the system state is called lock-in. Different regimes characterize the three phases:

Phase I is characterized by contingency and choice is unrestricted. Nonetheless, foregoing developments may have a slight narrowing impact, illustrating that history always matters (Teece et al. 1997; David 1994). Often a small event favours one of the possible solutions. If this solution enjoys positive feedback, a self-reinforcing dynamic may take place.

In phase II, these self-reinforcing mechanisms increasingly limit the scope of choice and thereby facilitate the evolution of an organizational path (Sydow et al. 2009; Stieglitz and Heine 2007; North 1990). In this context Arthur (1996) highlights the importance of increasing returns. Agents get (consciously or unconsciously) involved in a logic of repetition; they earn increasing returns through repetition as it is the case for example with economies of scale or network externalities. In a way they enjoy a rave of feedback and by doing so they unintentionally commit to path-building. The diminishing variety and the rising limitations of choices are collateral side effects of this process (Sydow et al. 2009).

With transition to phase III, the diminishing window of opportunity finally closes, leaving the organization strategically trapped in an unalterable state. The decision makers in the system are locked-in (Sydow et al. 2009). Empirical path dependence research has hallmarked the crucial elements that drive path emerging processes and finally lead to a potentially inefficient lock-in situation in phase III. At least four major types of self-reinforcing mechanisms can be distinguished: (1) coordination effects, (2) complementary effects, (3) learning effects, and (4) adaptive expectation effects (Sydow et al. 2009).

Since we are interested in path-dependent mechanisms in business-to-business relationships, we will concentrate on learning in and between organizations. Beyond, we aim to connect the often discussed and cited concept of absorptive capacity to path dependence theory.

Absorptive Capacity: Status Quo of the Seminal Conceptualizations

The transfer of knowledge is one of the main research areas in organizational learning and knowledge management (Easterby-Smith et al. 2008). Within this field the concept of absorptive capacity has gained wide attention describing the capability to incorporate and process valuable information (Cohen and Levinthal 1990). This capability is crucial for many organizations

in order maintain a high level of performance (Inkpen and Dinur 1998; Lane et al. 2006). Also a purposeful use of knowledge facilitates to achieve competitive advantages (Inkpen and Dinur 1998). For these reasons Lane et al. (2006) argue that the concept of absorptive capacity is “one of the most important constructs to emerge in organizational research in recent decades” (Lane et al. 2006: 833).

Regarding to Van den Bosch et al. (2003) absorptive capacity is a versatile concept, which can be applied to different theoretical and empirical problems and disciplines. The concept has been widely applied on different levels of analysis such as the organizational level (Cohen and Levinthal 1990; Boynton et al. 1994; Szulanski 1996; Veugelers 1997; Kim 1998), the inter-organizational level (Lane and Lubatkin 1998; Dyer and Singh 1998) and even the country level (Mowery and Oxley 1995; Keller 1996; Liu and White 1997). As our literature review reveals only little research has so far focused on the inter-organizational perspective. As one of the few authors in this field Lane and Lubatkin (1998) show that relatively similar knowledge bases and knowledge management systems within a dyadic alliance have a positive impact on the absorptive capacity of the partners. They should therefore be considered as important characteristics in inter-organizational learning.

We will proceed with summarizing the classical concept of absorptive capacity brought forward by Cohen and Levinthal (1990) as well as recent reconceptualizations with specific emphasis on inter-organizational and partner-specific absorptive capacity.

The Cohen and Levinthal (1990) model of absorptive capacity

A widely acknowledged definition of absorptive capacity was introduced by Cohen and Levinthal (1990). In their study, the authors examine businesses’ ability to innovate and define absorptive capacity as “an ability to recognize the value of new information, assimilate it, and apply it to commercial ends” (Cohen and Levinthal 1990: 128). Under this definition, absorptive capacity can be thought of as the ability to absorb new knowledge and is characterized by three key elements: (1) recognizing and assessing valuable new information, (2) assimilation of information classified as useful, and (3) use of this information for commercial purposes. The authors also point out that absorptive capacity arguably arises as a byproduct of a firm’s R&D investments (Cohen and Levinthal 1989; Cohen and Levinthal 1990; Cohen and Levinthal 1994).

Cohen and Levinthal (1990) also emphasize that the development of absorptive capacity depends on the amount of knowledge absorbed beforehand. In this context, the compatibility of old and new knowledge is important. Accordingly, absorptive capacity does not develop without any preconditions; instead, it is informed by earlier decisions. In addition, the formation of expectations and the behavior during future periods

are also affected by historical developments. Figure 2 summarizes the Cohen and Levinthal (1990) model of absorptive capacity.

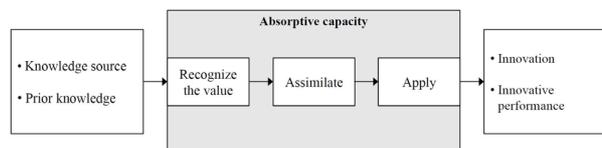


Figure 2: Model of Absorptive Capacity According to Cohen and Levinthal (1990)

In a later study, Cohen and Levinthal (1994) redefine absorptive capacity as a cluster of related abilities: “The capacity to ‘exploit’ outside knowledge is comprised of the set of closely related abilities to evaluate the technological and commercial potential of knowledge in a particular domain, assimilate it, and apply it to commercial ends. These abilities collectively constitute what we have termed a firm’s ‘absorptive capacity’” (Cohen and Levinthal 1994: 227).

The Zahra and George (2002) model of absorptive capacity

Zahra and George (2002) reconceptualize the construct of absorptive capacity as a dynamic capability (for a detailed discussion, see Teece et al. 1997), thereby defining the ability to absorb knowledge as a “set of organizational routines and processes by which firms acquire, assimilate, transform, and exploit knowledge to produce a dynamic organizational capability” (Zahra and George 2002: 186). The authors consequently assume that the abilities to acquire, assimilate, transform, and exploit knowledge build on one another and have a positive impact on the development of further capabilities. Table 1 defines the critical abilities within the scope of absorptive capacity.

Table 1: Definitions of Absorptive Capacity Abilities (Zahra and George 2002: 189-190)

Acquisition	<i>"[...] refers to a firm's capability to identify and acquire externally generated knowledge that is critical to its operations"</i>
Assimilation	<i>"[...] refers to the firm's routines and processes that allow it to analyze, process, and understand the information obtained from external sources"</i>
Transformation	<i>"[...] denotes a firm's capability to develop and refine the routines that facilitate combining existing knowledge and the newly acquired and assimilated knowledge"</i>
Exploitation	<i>"[...] is based on the routines that allow firms to refine, extend, and leverage existing competencies or to create new ones by incorporating acquired and transformed knowledge into its operations"</i>

Hence, Jansen et al. (2005) show empirically that a model based on these four separate factors is superior to models with fewer factors. In line with this conclusion, Todorova and Durisin (2007) recommend that researchers should choose models with four factors in future studies in order to use variables with high construct validity when testing their hypotheses.

The Todorova and Durisin (2007) model of absorptive capacity

Todorova and Durisin (2007) call for a systematic return to the key elements identified in the seminal article by Cohen and Levinthal (1990). First, they emphasize that 'recognize the value' should be seen as a central element. This component seems especially important as businesses often fail simply by not recognizing potentially relevant knowledge. Accordingly, new information is not grasped automatically. Rather, appreciation of new information is shaped in advance by existing structures. Consequently, absorptive capacity can be thought of as a necessary prerequisite for recognizing relevant new information.

Second, transformation of new information should be understood as an alternative process to assimilation and not as a process step following assimilation. Todorova and Durisin (2007) assume that external knowledge moves back and forth in a two-way process between the elements of assimilation and transformation before it can be successfully adopted and exploited within existing structures. In making this case, the authors draw on findings from learning theory: If external information is already largely in line with cognitive structures, new knowledge requires only minor changes before it is integrated into the knowledge base (assimilation). If there is no connection between the cognitive structures,

those structures first have to be transformed in order to absorb new knowledge.

In their model the two-way process of assimilation and transformation is formulated in the sense that external information can be absorbed even though it has no connection with the prior cognitive structures. For this to happen, however, the organization has to adapt its knowledge structures in the process. This is in line with Cohen and Levinthal's original model, in which absorption of new information depends on the amount or level of knowledge that has previously been absorbed. Finally, Todorova and Durisin incorporate positive feedback loops into their model, highlighting the dynamic character. This gives the largely one-dimensional model strength and mobility. At the same time, the authors point to feedback loops as a way of expanding the knowledge base during future periods. Figure 3 shows the dynamic model of absorptive capacity according to Todorova and Durisin (2007).

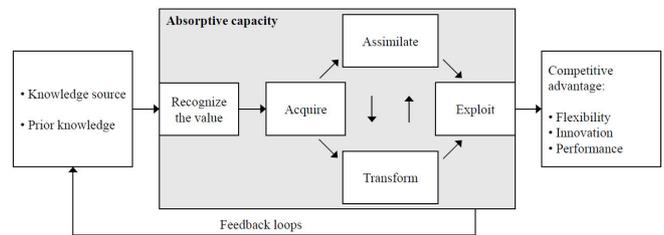


Figure 3: Model of Absorptive Capacity According to Todorova and Durisin (2007)

The Connection Between Absorptive Capacity and Path Dependence Theory

Lane et al. (2006) emphasize that the possibilities of the concept of absorptive capacity have not yet been fully exhausted and the concept still offers lots of potential for further research. This section aims to examine the connection between partner-specific absorptive capacity and the theory of path dependence.

As early as in their seminal article on absorptive capacity, Cohen and Levinthal point to the path-dependent character of this concept: "[...] the development of absorptive capacity, and, in turn, innovative performance are history- or path-dependent" (Cohen and Levinthal 1990: 128). The authors state that the path-dependent development of absorptive capacity depends on the existing knowledge base, pointing to findings from cognitive research and emphasizing that fully formed cognitive structures facilitate absorption of similar knowledge. Consequently, knowledge that has already been absorbed affects absorption of new knowledge while also affecting the formation of expectations in a self-reinforcing way.

Cohen and Levinthal (1990) also argue that the absorptive capacity that has grown historically in a certain area allows for more-efficient accumulation during subsequent periods. We can conclude that the following self-reinforcing process takes place: "absorptive capacity → learning → new absorptive

capacity” (Lane et al. 2006: 845). The corresponding absorptive capacity increases through repeated feedback loops, and increases the future capacity to absorb new related knowledge. Van den Bosch et al. (1999) also speak of “path dependency of absorptive capacity” (Van den Bosch et al. 1999: 554) and of a “path-dependence phenomenon of absorptive capacity” (Van den Bosch et al. 1999: 558). The self-reinforcing effect of absorptive capacity can thus be seen as a central feature of a path-dependent development (Cohen and Levinthal 1990; Van den Bosch et al. 1999). Therefore, the frequently mentioned link between absorptive capacity and its path-dependent character seems not surprising (Todorova and Durisin 2007; Lane et al. 2006; Zahra and George 2002; Lavie and Rosenkopf 2006; Lichtenthaler 2009).

Continuous investments in absorptive capacity also strengthen the ability to absorb knowledge, while at the same time shaping the path-like process. On the other hand, there is a risk that if investments are not made in other areas, critical fields of knowledge will be sidelined and not updated, creating a lock-out situation (Cohen and Levinthal 1990).

The publications on absorptive capacity reviewed above hint at its path-dependent character. This is also in line with recent findings presented by Mallach (2012). It is apparent that consistent development and expansion of absorptive capacity shapes its specific form and has a path-forming effect over time due to self-reinforcing feedback loops. The outcome is a learning path shaped by history (Van den Bosch et al. 1999). In the next section we will combine various arguments reviewed so far and propose an integrated path model of partner-specific absorptive capacity.

MODEL DEVELOPMENT: A PATH MODEL OF PARTNER-SPECIFIC ABSORPTIVE CAPACITY IN BUSINESS-TO-BUSINESS RELATIONSHIPS

Plinke (1997) points out that in business-to-business markets a seller firm increasingly focus on a single buyer firm and the specific buying firm’s perspective with regard to solving problems. This means that a seller firm faces the challenge of recognizing valuable knowledge held by its partner, assimilating that knowledge, and then utilizing it (Lane and Lubatkin 1998). One crucial factor here is partner-specific absorptive capacity, with the student-teacher concept being applied (Lane and Lubatkin 1998).

Subsequently, the influence of partner-specific absorptive capacity as a self-reinforcing mechanism is considered within the framework of the theory of path dependence. Also the question arises, whether working with a certain buyer firm for a longer period improves the seller firm’s partner-specific ability to absorb knowledge in this regard. From the perspective of path theory, it is especially interesting to consider whether, in these circumstances, bonding forces within the dyadic business relationship can result in lasting persistence or even a lock-in situation.

Conceiving of a Path Model of Partner-Specific Absorptive Capacity

For the development of a unifying path model, the three-phase model of path-dependent processes proposed by Sydow et al. (2009) serves as a reference system. The process of working through the three subsequent phases is a particular interest. After a description of the initial situation (phase I) the focus lies on the phase in which paths are formed (phase II). The dyadic buyer-seller relationship is conceptualized as a student-teacher relationship, with the seller firm in the position of the student and the buyer firm in that of the teacher. To reduce complexity, one-sided transfer of knowledge from the buyer (the teacher) to the seller (the student) is considered (Lane and Lubatkin 1998; Lane et al. 2006). In the process, the focus turns to the argument put forward by Plinke (1997) that the seller firm increasingly concentrates on a single business customer’s problem-solving perspective.

This paper also aims to show that business partners are over time subject to mechanisms that affect partner-specific absorptive capacity while also reinforcing it. It will be argued that established mechanisms of path theory such as coordination effects, complementary effects and adaptive expectation effects (Sydow et al. 2009) further reinforce the spiraling solidification of partner-specific absorptive capacity. As a result a learning path focusing on the specific relationship is formed by the student firm.

After many studies on absorptive capacity have mainly focused on success-oriented effects (e.g. Cohen and Levinthal 1990; Cohen and Levinthal 1994; Lichtenthaler 2009; Todorova and Durisin 2007; Zahra and George 2002), this paper takes a different approach, considering the potential dark side of learning processes. Thus far, this perspective is underrepresented in the literature.

Phase I: A Large Number of Potential Business Relationships and Learning Opportunities

The first phase is characterized by a large number of potential business partners with individual learning opportunities for the student firm. At this time, it is more or less unclear which companies will engage into a business relationship. It is likewise unclear for which partner the student firm will develop specific absorptive capacity abilities, and which knowledge will be absorbed in the future. Thus, the various possibilities are still largely unaligned and contingent.

Still, it is conceivable that a formative influence already exists during this early phase. Industry affiliation and standards or relevant market trends determine the field of action, for example, and select possible business partners.

If a business relationship ultimately forms based on a series of market transactions between a selling and a buying firm, specific knowledge is exchanged. We can

presume that the student firm has an interest in absorbing important knowledge quickly in order to recognize the teacher firm's needs in the best possible way and perform as desired. Development of partner-specific absorptive capacity is therefore desirable. Following this, a partner-oriented capability to absorb and process knowledge is concentrated in the student firm, while at the same time the effect of bonding with the teacher firm is intensified. This consciously or unconsciously solidifies the available space for new knowledge within the student firm, and a partner-related learning path evolves. This point in time can be viewed as a critical juncture, and thus as the transition to the second phase (Vergne and Durand 2010).

Phase II: Formation of a Learning Path Through Partner-Specific Absorptive Capacity

Phase II is the phase during which a path-dependent process is formed. Now that the student firm's focus is on the teacher firm's characteristic knowledge and the development of partner-specific absorptive capacity has begun, the development of a learning path featuring positive feedback arises within the student firm.

Since new knowledge can only be absorbed on the basis of existing knowledge, the knowledge that has been previously absorbed determines the development of absorptive capacity. This means that the existing knowledge base serves as the foundation for further knowledge, reinforcing itself step by step. This process makes it clear that absorptive capacity is, at bottom, a learning effect such as those described in path theory. In a study conducted by Mallach (2012), the author finds empirical support for the relationship between partner-specific absorptive capacity and learning effects and its path-dependent character in business-to-business relationships.

Next, we turn to interactions between the traditional self-reinforcing mechanisms of path theory and partner-specific absorptive capacity. While absorptive capacity is viewed as a special form of learning effects, its interactions with the other mechanisms involved, such as coordination effects, complementarity effects, and adaptive expectation effects (Sydow et al 2009) are of interest in terms of arriving at an integrated model. Sydow et al. (2009) already point out that several of these mechanisms are often at work at the same time. This paper will argue that coordination effects, complementarity effects, and adaptive expectation effects further enhance the self-reinforcing character of absorptive capacity.

Coordination effects and partner-specific absorptive capacity

Processes of coordination between interacting partners lie at the heart of the matter. The more actors within a system view rules as productive or adapt to accommodate routines or practices, the easier and more effective their interaction processes within the overall

system become (David 1994). As a result, a rule's overall attractiveness increases the more that rule is disseminated.

In her study on organizational routines, Knott (2003) points to the importance of routines as a mechanism for coordination. In this view, routines support the disruption-free progress of coordinated behavior. Dyer and Nobeoka (2000) similarly point out the importance of such coordination principles.

In their study of successful cooperation within strategic alliances, Dyer et al. (2001) propose that a dedicated strategic alliance function is responsible for success in these settings. The central tasks of this kind of function include optimized knowledge management, which in turn encompasses processes related to the articulation, documentation, and codification of knowledge, along with joint use of knowledge. In this way, the individual processes exert a direct influence on the expansion of the knowledge base. Tools, instruments, and templates that have been developed are also conducive to the further formation, definition, and solidification of routines of interaction (Dyer et al. 2001).

Dyer and Nobeoka (2000) also point out that shared knowledge transfer routines between interaction partners have a positive effect on inter-organizational learning. Consequently, interaction practices that are used regularly not only improve the transfer of knowledge between the interested parties, but also the formation of specialized knowledge. Routines relating to knowledge transfer also make it possible to store knowledge systematically and use it during subsequent periods (Dyer and Singh 1998; Dyer and Nobeoka 2000).

Van den Bosch et al. (1999) emphasize, in their analysis of absorptive capacity, that the ability to coordinate between group members has a positive effect on the absorption and processing of new knowledge. This includes deliberately constructed or indirectly developed processes of interaction between the involved parties at an organizational or inter-organizational level (Van den Bosch et al. 1999).

According to Dyer and Singh (1998), interaction routines between organizations promote recognition of important knowledge within the relationship. In this view, these routines have a direct influence on partner-specific absorptive capacity, supporting its development. Chang and Gotcher (2010) take a similar view. According to them, the functioning of partner-specific absorptive capacity can be improved by establishing routines related to interactions within the buyer-seller relationship.

Complementary effects and partner-specific absorptive capacity

Complementarity effects can be best described as synergistic effects. Synergies arise through repeated, mutually supportive interactions between separate but related resources, rules, and/or practices (Petermann 2010). Repeated interaction can give rise to utility that is greater overall than the sum of its parts (David 1994;

Petermann et al. 2012). Formation of core competencies is one possible consequence (for a detailed discussion, see Prahalad and Hamel 1990).

In the buyer-seller relationship, synergies can arise when student firms and teacher firms make deliberate efforts to share complementary knowledge. When the student firm's knowledge base grows, the company becomes able to recognize and absorb relevant new knowledge faster in subsequent periods. Partner-specific absorptive capacity increases as a result, influencing the future form and development of the characteristic learning path within the student firm.

Adaptive expectation effects and partner-specific absorptive capacity

Adaptive expectations confirm themselves step by step as part of a reciprocal process, thereby having a self-reinforcing effect. Preferences are not set from the start, instead forming as a result of the actors' expectations. The desire to belong socially or to be among the winners is a possible reason for adaptation. The more the actors' expectations associated with a certain behavior or an established approach are, the more attractive it is to adapt to accommodate these practices. Moreover, the behavior gains legitimacy as more and more actors align themselves to it. The reproduction and solidification of best practices is a good example (Sydow et al. 2009).

The Dacin et al. (2007) study on strategic alliances from an institutional perspective also refers to the role of legitimacy as a self-reinforcing mechanism. From the authors' standpoint, it is legitimate for individuals within an organization to copy stable structures and processes. The result, however, means that organizations can potentially become inflexible and unable to adapt rapidly enough to changing environmental influences (Dacin et al. 2007).

We assume that the business partners expect to share certain areas of their knowledge when they enter into a business relationship. It is also conceivable that after a series of transactions, certain practices of knowledge sharing (best practices) become established, and the interested parties might even no longer question these practices. Likewise, it is legitimate for the student firm in the buyer-seller relationship to absorb and process knowledge. Accordingly, development of partner-specific absorptive capacity is promoted.

Plinke (1997) points out that partners in a business relationship amass experience as a result of market transactions, and that new expectations form based on that experience. This process continues step by step, thereby reinforcing itself.

Phase III: Manifestation of a Potentially Inefficient Competency Trap Through One-Sided Absorption of Knowledge

The student firm's partner-specific absorptive capacity has reinforced itself and solidified during phase II as a result of repeated market transactions taking place

within the buyer-seller relationship. On the positive side, the student firm is now, in phase III, able to absorb new knowledge faster, provide innovative solutions for the teacher firm (e.g. Cohen and Levinthal 1990; Cohen and Levinthal 1994; Lichtenthaler 2009), and possibly even generate a competitive advantage for itself and its partner (e.g. Todorova and Durisin 2007; Zahra and George 2002).

The phase of path dependence (phase III) can be perceived as a one-sided learning path, which may be manifested in a potentially inefficient lock-in situation. Levinthal and March (1993) point to the shortsightedness of learning processes as being responsible for this potential inefficiency. When a longer-term perspective is neglected during the learning process, there is often the risk of a tendency to overlook failures. In this context, Zahra and George (2002) point out that path-dependent development of absorptive capacity determines not only an organization's success, but also its failure; they speak of a potential competency trap within learning paths. This paper will take up this approach and discuss the possibility of a learning or competency trap.

Abrupt changes in the external environment often challenge even established business firms. In dynamic environments, a dark side to partner-specific absorptive capacity may emerge. In that case necessary adaptation can – if at all possible – only take place gradually, and the risk of an inefficient competency or learning trap increases. In extreme cases, the risk of one-sided knowledge and the associated frame of reference, might result in neglecting more-efficient alternatives (Sydow et al. 2009; Levinthal and March 1993; Lei et al. 1996). Miller (1993) describes this development as “converting a formula for success into a path towards failure” (Miller 1993: 116).

If rigidity is so embedded that flexible realignment is no longer possible, the competency trap springs shut, manifesting itself in a potentially inefficient lock-in situation. Consequently, the direction in which the learning path has previously set out may be responsible for the potential failure of the student firm. Leonard-Barton (1992) describes this phenomenon as a capability-rigidity paradox, stating: “Core rigidities are the flip side of core capabilities” (Leonard-Barton 1992: 118).

Competency traps form as a result of various developments. In this context, Zahra and George (2002) mention three possible kinds of traps: the familiarity trap, the maturity trap, and the propinquity or nearness trap.

A familiarity trap results from excessive focus on refining and improving existing knowledge. As experience is amassed, absorptive capacity grows, and turning toward other alternatives seems not to be worthwhile. The argument of positive feedback can be raised again, as it can be the factor responsible for the solidification of knowledge. Thus, not only is exploration of alternative sources of knowledge

prevented, but cognitive structures also remain limited. If an actor does not succeed in absorbing different types of knowledge from different sources, the path-like course promotes rigidity and may even prevent a necessary paradigm shift (Ahuja and Lampert 2001; Zahra and George 2002).

In the maturity trap, the focus is on the need for dependable and predictable outcomes. Tapping into different kinds of knowledge from different sources fades into the background. Use of existing knowledge is also legitimate, although superior performance would be possible if outside sources of knowledge were used (Ahuja and Lampert 2001; Zahra and George 2002).

The propinquity or nearness trap describes a company's tendency to accumulate further knowledge in long-familiar areas. The tendency to give preference to the familiar is once again key, with other relevant areas of knowledge being ignored. This kind of trap becomes especially critical for seller firms when environmental conditions change, requiring completely new knowledge (Ahuja and Lampert 2001; Zahra and George 2002).

Often, competency and learning traps are also discussed in connection with exploitation and exploration of knowledge (e.g. Lavie and Rosenkopf 2006; Levinthal and March 1993; March 1991). According to March (1991), exploitation means the utilization and refinement of existing knowledge, while exploration means tapping into new knowledge and experimenting with unfamiliar, risky alternatives. In the process, firms should take care not only to exploit their existing knowledge base and the associated sources of knowledge, but to continue to explore new knowledge.

When environmental influences are stable, a strategy of exploitation is generally noncritical. But as positive experiences build on each other, the risk of path-dependent development also increases. While competence related to the existing activity rises, there is the risk that better alternatives are being neglected or not even perceived (March 1991). Long-term perspectives also might fade from view (Levinthal and March 1993). When environmental influences suddenly change, fresh knowledge is generally needed. If absorption of new knowledge does not take place, the benefits of an exploitation strategy can be reversed, suddenly turning into a competency trap (Koza and Lewin 1998; Lichtenthaler 2009).

A consistent alignment to the teacher firm's knowledge base becomes problematic if it means that the student firm ignores innovative developments in the field in which it operates, for example, or if other environmental influences (abruptly) change. If new knowledge is needed, the competency or learning trap may spring shut as the student firm's learning path becomes potentially inefficient. This is where the lock-in situation typical of paths is manifested, and there is a risk that the student firm will fail due to competition on the market (Zahra and George 2002). Because of the rather specific nature of the competency trap (the capacity of the student firm to recognize the value of new knowledge, acquire,

assimilate, exploit and transform it in a relational context), we introduce the term 'relational absorptive capacity trap'.

In summary, one-sided concentration by the student firm on partner-specific absorptive capacity, and simultaneously on the teacher firm's knowledge base, pushes aside exploration of new areas of knowledge lying outside the business relationship. This leads to a potentially inefficient relational absorptive capacity trap within the student firm, a trap that may spring shut if environmental conditions change abruptly, manifesting itself in a lock-in situation.

The student firm might face similar difficulties if the teacher firm disappears from the market. There is also the risk that the student firm cannot move off its existing learning path fast enough to realign itself on the market. The internal connection between the student firm and the teacher firm is also solidified by investments in partner-specific absorptive capacity. These relationship-oriented costs are also known as switching costs and lost when the partnership ends. Because of their highly specific character, these investments cannot be reused elsewhere (Geiger et al. 2012).

The bonding forces that arise in these situations make it seem that the further success of a student firm depends on its relationship with its teacher firm. According to Narasimhan et al. (2009), the lock-in situation in which the student firm finds itself is manifested here in the form of a dependent relationship. Therefore, we posit that one-sided concentration by the student firm on partner-specific absorptive capacity, and simultaneously on the teacher firm's knowledge base, increases the forces binding the two companies. Also, it causes the student firm to be dependent on the teacher firm in order to continue to exist on the market.

Towards a Path Model of Partner-Specific Absorptive Capacity

The foregoing discussion has shown how partner-specific absorptive capacity can be viewed as a self-reinforcing mechanism within the framework of the theory of path dependence. To this end, the paper has argued based strictly on the reference system of the three-phase model of path-dependent processes put forward by Sydow et al. (2009). The content of the individual phases in this model was applied to the buyer-seller relationship in a business-to-business marketing context. This section will now summarize the insights gleaned through this process in the form of a path model showing the position of the seller respectively the student firm, as the case may be (see Figure 4). Particular attention is paid to ensuring that the perspective of time is taken into account across the entire model. This means, first, that the model accommodates a path-dependent process (Sydow et al. 2009); and second, the definition of a business-to-business relationship as a non-accidental sequence of market transactions (Kleinaltenkamp and Ehret 2006).

Phase I shows a large number of potential business partners for the student firm, with individual learning opportunities. In general, the various possibilities are still unaligned and contingent at this point. After a series of market transactions, a business-to-business relationship ultimately forms, and specific knowledge is normally transferred from the teacher firm to the student firm. To absorb and process this knowledge, the student firm develops partner-specific absorptive capacity, and a characteristic learning path forms.

Phase II shows the development of the partner-oriented learning path in the student firm, featuring positive feedback and is further enhanced by a series of feedback loops. The crucial point here is the self-reinforcing effect of partner-specific absorptive capacity. The relationship-oriented competency of knowledge absorption is concentrated within the student firm, and the effect of being bound to the teacher firm increases. This process is promoted by coordination effects, complementarity effects, and adaptive expectation effects.

The concept of partner-specific absorptive capacity as treated herein is oriented toward the dynamic model proposed by Todorova and Durisin (2007). This model was chosen because it has been empirically tested that a model with four separate factors (acquisition, assimilation, transformation, and exploitation) is preferable to one with only two factors (Jansen et al. 2005). Beyond that, this approach is interesting in that new knowledge moves back and forth within a two-way process of assimilation and transformation before it can be successfully used for commercial purposes (Todorova and Durisin 2007). The integration of the learning process-oriented perspective on absorptive capacity put forward by Lane et al. (2006) – with the learning steps of exploratory learning, transformative learning, and exploitative learning – represents a valuable addition. In the developed path model, exploratory learning is applied to the student-teacher concept and understood as absorption of new knowledge within the dyadic business-to-business partnership.

Phase III points to possible competency or learning traps in one-sided learning paths. Due to the rather specific nature of these traps, we introduced the term relational absorptive capacity trap. It should be assumed that in the case of changed environmental influences, for example, the relational absorptive capacity trap will spring shut, solidifying into a potentially inefficient learning path within the student firm. This marks the ambivalent nature of partner-specific absorptive capacity.

Based on our literature review and theoretical considerations, we suggest a path model of partner-specific absorptive capacity as presented in Figure 4.

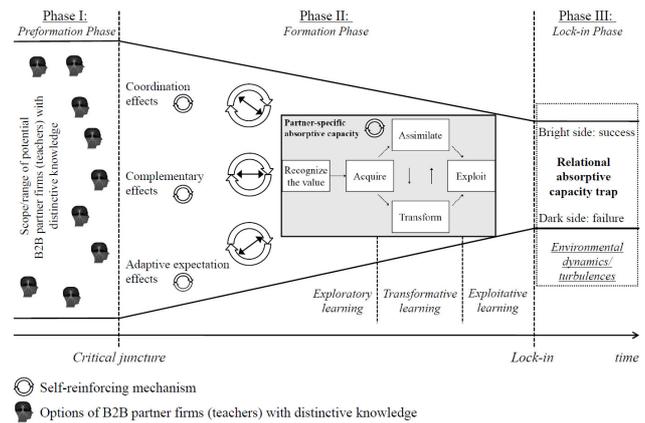


Figure 4: Path Model of Partner-Specific Absorptive Capacity in B2B Relationships

FROM AN INTEGRATED PATH MODEL OF ABSORPTIVE CAPACITY IN B2B RELATIONSHIPS TOWARDS A SIMULATION MODEL

We suggest that partner-specific absorptive capacity lends itself to being thought of as a self-reinforcing mechanism within the framework of the theory of path dependence. Moreover, we argue that coordination effects, complementarity effects, and adaptive expectation effects reinforce the spiraling, self-reinforcing character of partner-specific absorptive capacity, and thus may contribute to the further development and definition of a characteristic learning path. This would increase the effect of binding the student firm to the teacher firm, and there is a risk of a potentially inefficient competency or learning trap that could spring shut under certain conditions. To specify the competency or learning trap, we introduced the term relational absorptive capacity trap. Accordingly, the path model developed here considers the often-disregarded dark side of learning processes.

The theoretical model we derived above offers potential for future simulation research. A computer simulation study could be conducted to further explore the model, understand the interdependencies and mechanisms at hand and identify the critical factors involved in the formation of a learning path (Vergne and Durand 2010). Computer simulations are nowadays a well established method in social sciences (Gilbert and Troitzsch 2005; Harrison et al. 2007) and scholars have found convincing answers to the arguments of critics of simulation modeling (Waldherr and Wijermans 2013). Also, computer simulations have recently been successfully applied in the field of marketing (e.g. Rand and Rust 2011) and organizational learning (Siggelkow and Rivkin 2006). In the case of a byersellerrelationship we find several aspects that hint to an agent based model (ABM) for further investigation. First, we have a dyadic relationship that develops gradually between two or more actors. Second, we showed that the process unfolding over time is governed by one or more self-reinforcing mechanisms on different

levels. Such complex multi-level processes often cannot be captured by analytical means only and numeric approaches are needed to uncover the long-term behavior of the system (Petermann et al. 2012). To sum up: Simulation modeling is particularly fruitful when complex, longitudinal and multi-level processes are at hand (Davis et al. 2007). This is generally true for path-dependent processes (Schreyögg and Sydow 2011) and specifically for the relationships governed by partner-specific absorptive capacity. Agent based modeling and simulation research in this area have not been conducted yet and could contribute to shedding more light on the nature of path-dependent partner-specific absorptive capacity as an interesting phenomenon in the development of business-to-business relationships.

REFERENCES

- Ahuja, G. and C.M. Lampert. 2001. "Entrepreneurship in the Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Inventions". In *Strategic Management Journal*, 22 (6/7), 521-543.
- Arthur, W.B. 1989. "Competing Technologies, Increasing Returns, and Lock-in by Historical Events". In *The Economic Journal*, 99 (394), 116-131.
- Arthur, W.B. 1996. "Increasing Returns and the New World of Business". In *Harvard Business Review*, 74 (4), 100-109.
- Boynton, A.C.; R.W. Zmud; and G.C. Jacobs. 1994. "The Influence of IT Management Practice on IT Use in Large Organizations". In *MIS Quarterly*, 18 (3), 299-320.
- Brennan, R.; L. Canning; and R. McDowell. 2007. *Business-to-Business Marketing*. Los Angeles, CA et al.: Sage.
- Chang, K.-H. and D.F. Gotcher. 2010. "Conflict-Coordination Learning in Marketing Channel Relationships: The Distributer View". In *Industrial Marketing Management*, 39 (2), 287-297.
- Chen, Y.-S.; M.-J.J. Lin; and C.-H. Chang. 2009. "The Positive Effects of Relationship Learning and Absorptive Capacity on Innovation Performance and Competitive Advantage in Industrial Markets". In *Industrial Marketing Management*, 38 (2), 152-158.
- Cohen, W.M. and D.A. Levinthal. 1989. "Innovation and Learning: The Two Faces of R&D". In *The Economic Journal*, 99 (397), 569-596.
- Cohen, W.M. and D.A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation". In *Administrative Science Quarterly*, 35 (1), 128-152.
- Cohen, W.M. and D.A. Levinthal. 1994. "Fortune Favors the Prepared Firm". In *Management Science*, 40 (2), 227-251.
- Dacin, M.T.; C. Oliver; and J.-P. Roy. 2007. "The Legitimacy of Strategic Alliances: An Institutional Perspective". In *Strategic Management Journal*, 28 (2), 169-187.
- David, P.A. 1985. "Clio and the Economics of QWERTY". In *American Economic Review*, 75 (2), 332-337.
- David, P.A. 1994. "Why are Institutions the 'Carriers of History'? Path Dependence and the Evolution of Conventions, Organizations and Institutions". In *Structural Change and Economic Dynamics*, 5 (2), 205-220.
- David, P.A. 2001. "Path Dependence, its Critics and the Quest for 'Historical Economics'". In P. Garrouste and S. Ioannides (Eds.). *Evolution and Path Dependence in Economic Ideas. Past and Present*. Cheltenham, UK / Northampton, MA: Elgar, 15-40.
- Davis, J.P.; K.M. Eisenhardt; and C.B. Bingham. 2007. "Developing Theory Through Simulation Methods". In *Academy of Management Review*, 32 (2), 480-499.
- Dyer, J.H. and H. Singh. 1998. "The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage". In *Academy of Management Review*, 23 (4), 660-679.
- Dyer, J.H. and K. Nobeoka. 2000. "Creating and Managing a High-Performance Knowledge-Sharing Network: The Toyota Case". In *Strategic Management Journal*, 21 (3), 345-367.
- Dyer, J.H.; P. Kale; and H. Singh. 2001. "How To Make Strategic Alliances Work". In *MIT Sloan Management Review*, 42 (4), 37-43.
- Easterby-Smith, M.; M.A. Lyles; and E.W.K. Tsang. 2008. "Inter-Organizational Knowledge Transfer: Current Themes and Future Prospects". In *Journal of Management Studies*. 45 (4), 677-690.
- Geiger, I.; A. Durand; S. Saab; M. Kleinaltenkamp; R. Baxterand; and Y Lee. 2012. "The Bonding Effects of Relationship Value and Switching Costs in Industrial Buyer-Seller Relationships: An Investigation Into Role Differences". In *Industrial Marketing Management*, 41 (1), 82-93.
- Gilbert, N. and K. Troitzsch. 2005. *Developing Theory through Simulation Methods*. Berkshire: Open University Press.
- Gulati, R. and M. Sytch. 2007. "Dependence Asymmetry and Joint Dependence in Interorganizational Relationships: Effects of Embeddedness on a Manufacturer's Performance in Procurement Relationships". In *Administrative Science Quarterly*, 52 (1), 32-69.
- Harrison, J.R.; Z. Lin; G.R. Carroll; and K.M. Carley. 2007. "Simulation Modeling in Organizational and Management Research". In *Academy of Management Review*, 32 (4), 1229-1245.
- Inkpen, A.C. and A. Dinur. 1998. "Knowledge Management Processes and International Joint Ventures". In *Organization Science*, 9 (4), 454-468.
- Jansen, J.J.P.; F.A.J. Van den Bosch; and H.W. Volberda. 2005. "Managing Potential and Realized Absorptive Capacity: How Do Organizational Antecedents Matter?". In *Academy of Management Journal*, 48 (6), 999-1015.
- Keller, W. 1996. "Absorptive Capacity: On the Creation and Acquisition of Technology in Development". In *Journal of Development Economics*, 49 (1), 199-210.
- Kim, L. 1998. "Crisis Construction and Organizational Learning: Capability Building in Catching-Up at Hyundai Motor". In *Organization Science*, 9 (4), 506-521.
- Kleinaltenkamp, M. 1997. "Kooperationen mit Kunden". In M. Kleinaltenkamp and W. Plinke (Eds.). *Geschäftsbeziehungsmanagement*. Berlin et al.: Springer, 219-276.
- Kleinaltenkamp, M. and M. Ehret. 2006. "The Value Added by Specific Investments: A Framework for Managing Relationships in the Context of Value Networks". In *Journal of Business & Industrial Marketing*, 21 (2), 65-71.
- Knott, A.M. 2003. "The Organizational Routines Factor Market Paradox". In *Strategic Management Journal*, 24 (10), 929-943.
- Koza, M.P. and A.Y. Lewin. 1998. "The Co-Evolution of Strategic Alliances". In *Organization Science*, 9 (3), 255-264.

- Lane, P.J. and M. Lubatkin. 1998. "Relative Absorptive Capacity and Interorganizational Learning". In *Strategic Management Journal*, 19 (5), 461-477.
- Lane, P.J.; B.R. Koka; and S. Pathak. 2006. "The Reification of Absorptive Capacity: A Critical Review and Rejuvenation of the Construct". In *Academy of Management Review*, 31 (4), 833-863.
- Lavie, D. and L. Rosenkopf. 2006. "Balancing Exploration and Exploitation in Alliance Formation". In *Academy of Management Journal*, 49 (4), 797-818.
- Lei, D.; M.A. Hitt; and R. Bettis. 1996. "Dynamic Core Competences Through Meta-Learning and Strategic Context". In *Journal of Management*, 22 (4), 549-569.
- Leonard-Barton, D. 1992. "Core Capabilities and Core Rigidities: A Paradox in Managing New Product Development". In *Strategic Management Journal*, 13 (Special Issue), 111-125.
- Levinthal, D.A. and J.G. March. 1993. "The Myopia of Learning". In *Strategic Management Journal*, 14 (Special Issue), 95-112.
- Lichtenthaler, U. 2009. "Absorptive Capacity, Environmental Turbulence, and the Complementarity of Organizational Learning Processes". In *Academy of Management Journal*, 52 (4), 822-846.
- Liu, X. and R.S. White. 1997. "The Relative Contributions of Foreign Technology and Domestic Inputs to Innovation in Chinese Manufacturing Industries". In *Technovation*, 17 (3), 119-125.
- Mallach, R.J. 2012. *Pfadabhängigkeit in Geschäftsbeziehungen*. Wiesbaden: Springer Gabler.
- March, J.G. 1991. "Exploration and Exploitation in Organizational Learning". In *Organization Science*, 2 (1), 71-87.
- Miller, D. 1993. "The Architecture of Simplicity". In *Academy of Management Review*, 18 (1), 116-138.
- Mowery, D.C. and J.E. Oxley. 1995. "Inward Technology Transfer and Competitiveness: The Role of National Innovation Systems". In *Cambridge Journal of Economics*, 19 (1), 67-93.
- Narasimhan, R.; A. Nair; D.A. Griffith; J.S. Arlbjörn; and E. Bendoly. 2009. "Lock-In Situations in Supply Chains: A Social Exchange Theoretic Study of Sourcing Arrangements in Buyer-Supplier Relationships". In *Journal of Operations Management*, 27 (5), 374-389.
- North, D.C. 1990. *Institutions, Institutional Change and Economic Performance*, Cambridge, UK et al.: Cambridge University Press.
- Petermann, A. 2010. *Pfadabhängigkeit und Hierarchie: Zur Durchsetzungskraft von selbstverstärkenden Effekten in hierarchischen Organisationen*, Berlin.
- Petermann, A.; S. Klaußner; and N. Senf. 2012. "Organizational Path Dependence: The Prevalence of Positive Feedback Economics in Hierarchical Organizations". In K.G. Troitzsch; M. Möhring; and U. Lotzmann (Eds.). *Proceedings 26th European Conference on Modeling and Simulation*, Koblenz, 721-730.
- Pierson, P. 2000. "Increasing Returns, Path Dependence, and the Study of Politics". In *American Political Science Review* 94 (2), 251-267.
- Plinke, W. 1997. "Grundlagen des Geschäftsbeziehungsmanagements". In M. Kleinaltenkamp and W. Plinke (Eds.). *Geschäftsbeziehungsmanagement*, Berlin et al., Springer, 1-62.
- Plinke, W. 2000. "Grundlagen des Marktprozesses". In M. Kleinaltenkamp and W. Plinke (Eds.). *Technischer Vertrieb: Grundlagen des Business-to-Business Marketing*, Berlin, Springer, 1-109.
- Prahalad, C.K. and G. Hamel. 1990. "The Core Competence of the Corporation". In *Harvard Business Review*, 68 (3), 79-91.
- Rand, W. and R.T. Rust. 2001. "Agent-Based Modeling in Marketing: Guidelines for Rigor". In *International Journal of Research in Marketing*, 28 (3), 181-193.
- Schreyögg, G. and J. Sydow. 2011. "Organizational Path Dependence: A Process View". In *Organization Studies*, 32 (3), 321-335.
- Siggelkow, N. and J.W. Rivkin. 2006. "When Exploration Backfires: Unintended Consequences of Multilevel Organizational Search". In *Academy of Management Journal*, 49 (4), 779-795.
- Sydow, J.; G. Schreyögg; and J. Koch. 2009. "Organizational Path Dependence: Opening the Black Box". In *Academy of Management Review*, 34 (4), 689-709.
- Stieglitz, N. and K. Heine. 2007. "Innovations and the Role of Complementaries in a Strategic Theory of the Firm". In *Strategic Management Journal*, 28 (1), 1-15.
- Szulanski, G. 1996. "Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm". In *Strategic Management Journal*, 17 (Special Issue), 27-43.
- Teece, D.J.; G. Pisano; and A. Shuen. 1997. "Dynamic Capabilities and Strategic Management". In *Strategic Management Journal*, 18 (7), 509-533.
- Todorova, G. and B. Durisin. 2007. "Absorptive Capacity: Valuing a Reconceptualization". In *Academy of Management Review*, 32 (3), 774-786.
- Van den Bosch, F.A.J.; H.W. Volberda; and M. De Boer. 1999. "Coevolution of Firm Absorptive Capacity and Knowledge Environment: Organizational Forms and Combinative Capabilities". In *Organization Science*, 10 (5), 551-568.
- Van den Bosch, F.A.J.; R. Van Wijk; and H.W. Volberda. 2003. "Absorptive Capacity: Antecedents, Models, and Outcomes". In M. Easterby-Smith and M. Lyles (Eds.). *The Blackwell Handbook of Organizational Learning and Knowledge Management*. Oxford, UK: Blackwell, 278-302.
- Vergne, J.-P. and R. Durand. 2010. "The Missing Link Between the Theory and Empirics of Path Dependence: Conceptual Clarification, Testability Issue, and Methodological Implications". In *Journal of Management Studies*, 47 (4), 736-759.
- Veugelers, R. 1997. "Internal R&D Expenditures and External Technology Sourcing". In *Research Policy*, 26 (3), 303-315.
- Von Hippel, E. 1988. *The Sources of Innovation*. NY et al.: Oxford University Press.
- Waldherr, A. and N. Wijermans. 2013. "Communicating Social Simulation Models to Sceptical Minds". In *Journal of Artificial Societies and Social Simulation*, 16 (4), 13.
- Zahra, S.A. and G. George. 2002. "Absorptive Capacity: A Review, Reconceptualization, and Extension". In *Academy of Management Review*, 27 (2), 185-203.

Tobias Grossmann, M.Sc. is a Research Associate in the Marketing-Department of the Freie Universität Berlin. Business-to-Business Marketing is his main research and teaching area. His research focuses on path-



dependence theory, fairness in business-to-business relationships, industrial negotiation and contract management. His research projects have been presented at several international academic conferences. As the Program Manager for the “China-Europe Executive Master of Business Marketing” postgraduate program, he has been involved in coordinating the development and commercialization of the program. In addition, he is responsible for sales, marketing and public relations.



Dr. Arne Petermann was MBA Program Director at the Berlin University for Professional Studies from 2011 until 2013. As a visiting scholar, he is currently teaching organization science and scientific simulation methods in the PhD-program at the School for Business and Economics at Freie Universität Berlin. He is founding President and CEO of Linara GmbH, an international HR agency specialized in the transnational European healthcare sector. His research focuses on organization science, path-dependence theory, entrepreneurship, and social simulation, especially agent-based modeling. His research results have been presented at the leading international science conferences in his field, including the Academy of Management and the American Marketing Association Educators Conference where, moreover, his work was recognized with a best paper award. His work is published in books and peer-reviewed journals.

SIMULATION OF SCHEDULING AND COST EFFECTIVENESS OF NURSES USING DOMAIN TRANSFORMATION METHOD

Geetha Baskaran
University of Nottingham Malaysia Campus
Jalan Broga , 43500 Semenyih, Malaysia
E-mail: Geetha.Baskaran@nottingham.edu.my

Andrzej Bargiela
University of Nottingham, Jubilee Campus &
Krakow Technical University. Krakow, Poland.
E-mail: abb@cs.nott.ac.uk

Rong Qu
School of Computer Science
University of Nottingham, Jubilee Campus
Nottingham NG8 1BB, UK.
E-mail: rxq@cs.nott.ac.uk

KEYWORDS

Domain Transformation, Nurse Scheduling, Simulation, Granular Computing, Integer Programming

ABSTRACT

Nurse scheduling is a complex combinatorial optimization problem. With increasing healthcare costs, and a shortage of trained staff it is becoming increasingly important for hospital management to make good operational decisions. A major element of hospital expenditure is staff cost. In order to help Kajang Hospital to make decisions about staffing and work scheduling, a simulation model was created to analyse the impact of alternate work schedules and investigate the optimum balance between the staffing levels of the ward and the ability to achieve good quality schedules. In this paper, we extend our novel approach to solve the nurse scheduling problem by transforming it through Information Granulation. This approach satisfies the rules of a typical hospital environment based on a real data set benchmark problem from Kajang Hospital.

Generating good work schedules has a great influence on nurses' working condition which is strongly related to the level of a quality health care. Domain transformation is an approach to solving complex problems that relies on well-justified simplification of the original problem. Solution of such a simplified problem and subsequent refinement of this solution to compensate for the simplifications introduced in the first step. Compared to conventional methods, our approach involves judicious grouping (information granulation) of shifts types' that transforms the original problem into a smaller solution domain. Later these schedules from the smaller problem domain are converted back into the original problem domain by taking into account the constraints that could not be represented in the smaller domain. An Integer Programming (IP) is formulated to solve the transformed scheduling problem by expending the branch and bound algorithm. We have used the GNU Octave, open source mathematical modelling and simulation software for Windows to solve this problem.

Results from simulations on real data problem sets for a typical hospital in Malaysia shows that this algorithm facilitated computation of feasible schedules in a short time with non-critical constraints being satisfied to a large degree. The resulting solutions facilitated cost benefit analysis of different staffing levels.

INTRODUCTION

Nurse scheduling problem, popularly known as (NSP), is the task of assigning an appropriate and efficient work regime for nurses in both private and government hospitals. Scheduling in an organization is very important to ensure the process of managing the company is effective and efficient. According to (Henderson V.A,1939) "The unique function of the nurse is to assist the individual, sick or well, in the performance of those activities contributing to the health or its recovery that he would perform unaided if he had the necessary strength, will or knowledge". Therefore, in order for the nurses to perform their job well, they need to be organised and this resulted in the formation of nurse scheduling. Nurse scheduling is often done manually, takes too much time and seldom show best quality results (Bouarab et al, 2010).

This may seem like an easy task but in reality it's something that requires a lot of effort and is very time consuming. Several requirements must be taken into account such as a minimal allocation of a ward, legal regulations and personal needs of the nurse (Abdennadher & Schlenker, 1999). In NSP, there are two types of constraint. They are hard constraint and soft constraint. Hard constraint is a rule that need to be encountered at all times or else the schedule is counted to be infeasible and not accepted. Soft constraint is operated to estimate the quality of the solution. So, soft constraint is not necessary but is required to be fulfilled as much as possible. Nevertheless, to get a schedule that encounters all the hard constraints it is frequently required breaking some of the soft rules. A weight is allocated for each soft constraint to reflect its worth. The objective of nurse scheduling is to find a schedule that

satisfies all hard constraints and minimises the degree to which the soft constraints are violated.

In this study, we present an alternative way of tackling a large, real world nurse scheduling problem by using integer programming (IP). With this approach, the hospital is supplied with detailed information about the schedule, which they can use to make the selection objectively. We use the domain transformation method introduced in [Baskaran et al. 2009, 2012,2013] as a practical illustration of the information granulation methodology [Bargiela et al. 2002, 2008] to generate multiple feasible low cost rosters, which are evaluated with simulation. Domain transformation is an approach to solving complex problems that relies on well-justified simplification of the original problem. We subdivided the problem into smaller subproblems in a systematic way and capable to reproduce the result. This approach is able to conquer solution easily by avoiding random search. Conversely, in other methods, some failed to reproduce results, and produce inconsistent performance, some works best on some datasets but failed to repeat the good characteristics on other datasets. The previous state-of-the-art never used Information Granulation (Domain Transformation Approach-DTA), thus dealing with a lot of cross referencing and checking of data. We have also approached the problem using the demand simulation to check the cost effective scheduling of nurses using the domain transformation method. In this paper, we discuss the process and results from these method.

NURSE SCHEDULING PROBLEM AT KAJANG HOSPITAL

The scheduling problem presented in this paper has been studied for three ward in a large Malaysian hospital. The problem is based on the situation of coronary care unit (CCU), medical ward and male ward in Kajang Hospital. We outline the following characteristics.

1. We have to adhere to Malaysia national laws, and the collective labor agreements enforced in Malaysian hospitals.
2. The requests of the personnel are very important, and should be met as much as possible; the soft constraints we use are those that, in our experience, represent the situation in Kajang hospital.
3. It is not necessary to consider qualifications, as all personnel are highly qualified. However, the specialized nurses are required to oversee all the tasks in each shift.

There are ten specialized nurses and eighteen normal nurses in Kajang Hospital. All the nurses are full time and have a contract of 40 hours per week. There are 28 scheduling problems involves assigning a certain number of different types of shifts as illustrated in Table

1 which satisfies the daily coverage requirements for these shift types.

Each of the shift types cover different number of hours including one hour of rest time. Early and Late shift covers 7 hours, day-shift covers 9 hours, and night shift covers 10 hours. The scheduling period practised in the hospital is 2 weeks. There are few types of rest days practised in this hospital. They are Sleep Day (SD), Day Off (DO), Public Holiday (PH), Annual Leave (AL), and Emergency Leave (EL).

Table 1: Shift Types and Demand during a week

Shift type	Start time	End Time	Demands						
			M	T	W	T	F	S	S
Early	07:00	14:00	6	6	6	6	6	6	6
Day	08:00	17:00	1	1	1	1	1	1	1
Late	14:00	21:00	6	6	6	6	6	6	6
Night	21:00	07:00	3	3	3	3	3	3	3

Early = E Day = D Late = L Night= N

Hard Constraints

The hard constraints listed below must be met in any conditions otherwise the schedule is considered to be infeasible and unacceptable. The hard constraints for NRP at Kajang Hospital are:

- HC1: Demands need to be fulfilled
- HC2: For each day, 1 nurse may start only one shift.
- HC3: One of the employee requires to perform only the Office Hour shift per day
- HC4: At least one skilled nurse must be scheduled to each shift.
- HC5: The number of consecutive shifts (night) is at most 3.
- HC6: The number of consecutive shifts (workdays) is at most 6.
- HC7: Following a series of 3 consecutive night shifts, a 48 hours rest is required.
- HC8: Following a series of 6 consecutive day shifts, a 24 hours rest is required.
- HC9: The maximum number of night shifts is 3 per period of 2 consecutive weeks.

Soft Constraints

Ideally these constraints should be satisfied as much as possible. However, in real world circumstances, it is usually necessary to violate some of these soft constraints. Depending on how strongly these soft constraints are desired (especially in comparison to other soft constraints), a weight is assigned to each of them. Soft constraints replicate the general preferences of the nurses and hospital's requirements at Kajang Hospital. The weights of soft constraints in the Kajang Hospital are described in Table 2.

Table 2: Soft constraints and their weights

Soft Constraints		Weights
SC1	Avoid sequence of shifts with length of 1 for all nurses.	1000
SC2	The rest after a series of <i>morning</i> or <i>evening</i> shifts is at least 2 days.	100
SC3	The number of shifts is within the range [4, 6] per week	10
SC4	The length of a series of shifts should be within the range of [4, 6].	10
SC5	Day on/off request. Requests by nurses to work or not to work on specific days of the week should be respected, otherwise solution quality is compromised	10
SC6	Shift On/Off Request. Similar to the previous but now for the specific shifts on certain days.	10
SC7	For all nurses, the length of a series of <i>morning</i> shifts should be within the range [1, 4]. It could be within another series of shifts.	10
SC8	For all nurses the length of a series of <i>evening</i> shifts should be within the range of [1, 4]. It could be within another series of shifts.	10
SC9a	A <i>morning</i> shift after the <i>office hour</i> shift should be avoided.	5
SC9b	An <i>evening</i> shift after the <i>office hour</i> shift should be avoided.	5
SC10	An evening shift after the day off that follows by with night shift	1

Objective Function

The objective function aims to minimize the total penalty of the soft constraints violation. A penalty weight is given for each soft constraint based on the importance of that constraint. So, this penalty weighting is simply a number. The higher the weight, the more strongly desired the satisfaction of this constraint is. The penalty of a feasible schedule is the sum of the weights of all the violations of soft constraints in the schedule. One key concern regarding setting the weights of constraints in NSPs is that, there are no standard weights to be given for each soft constraint. This is based on the wide range of constraints that are diverse from one hospital to another. Table 2 presents the weight of each soft constraint. As a channel, the weights could be described as follows:

Weight 1000 : The constraint should not be violated unless absolutely necessary.

Weight 100 : The constraint is strongly desired.

Weight 10 : The constraint is preferred but not critical.

Weight 5: The constraint is favoured but not crucial.

Weight 1 : Try and obey this constraint if possible but it is not essential.

PROPOSED SOLUTION OF SIMPLIFIED PLAN FOR SIMULATION

The simulation model describes the functioning of the main processes in a nurse scheduling problem. In our (figure 1), we have not only performed an efficient scheduling simulation but also presented a cost effective schedule by executing the demand simulation.

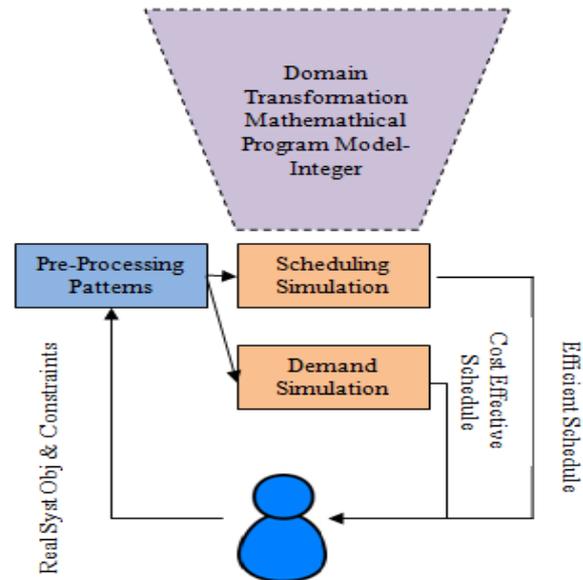


Figure 1: Design of the simulation model

Interactive scheduling is facilitated in our approach. The novelty in our research in contrary with other solutions. Our solution is focused on solving complex problems that relies on well-justified simplification of the original problem. We subdivided the problem into smaller subproblems in a systematic way and capable to reproduce the result. This identification of interactive scheduling is dynamic . It is thus fully independent user oriented and compatible with the new human centred computing paradigm. It is important to have easily understandable results in both domains. Another benefit in this simulation model is that the domain transformation can reduce computational complexity and therefore reduce the computational time. Besides, it also reduces the cross referencing over the detailed swapping of shifts for individual nurses. The goal of this scheduling simulation is to test how the different schedules perform when, for instance, the workload or capacity copes with uncertainty. In order to retrieve meaningful results, the simulation was tested intensively with a range of different parameters. This was then discussed with the real system materon to identify a number of service criteria in coordination with the hospital. If the first results indicate that the schedule do not meet the goals set in the simulation model, it can be necessary to adjust or add some of the constraints in the

mathematical programming model. They can choose which schedule to implement, since they have been provided with all the information they need concerning the different steps. Through this simulation model, the schedule obtained will not make any difference in terms of the different order of processing. The schedule is the same when we change the order of individual patterns or nurses.

Balancing The Cost Of Soft Constraints And The Staff Cost

Nurse scheduling is inextricably linked with determining total number of nurses. Most of healthcare systems are under pressure to control costs while trying to provide high levels of service. This is a difficult balance to strike. Establishing nurse scheduling model is a delicate balance between enhancing patient safety and provider productivity while also optimizing organizational costs. With demands to improve patients' clinical outcomes and decrease the escalating costs of inpatient care, nurses are focusing on how nurses spend their time rather than just raising staffing levels to positively impact patient outcomes. Because nursing wages constitute a high proportion of a hospital's budget, understanding the costs of number of nurses required is critical to manage them. Having a small number of nurses may impact quality of care while employing a large number of nurses and not utilizing their contractual hours is clearly wasteful. In our approach, we are balancing these concerns by combining the cost for the underutilization of nurses with the costs of violation of soft constraints into a single performance index. This process is discussed in (Baskaran, G., 2013). Hence, based on this demand simulation, we have produced numerical simulation experiments for this current nurse scheduling problem. In this problem we assume that the following represents well the notional cost of underemployed staff:

$$CU = U * 0.2 \quad (1)$$

where:

CU = Cost of under- utilization

DOMAIN TRANSFORMATION USING INTEGER PROGRAMMING

The domain transformation approach introduced in [Baskaran et al. 2012, 2013] departs from the convention of direct exploration of the space of schedules, as described in the preface section. We observe that the three shifts (e, d, l) are subject to identical soft constraints. Consequently, the first overview of the scheduling problem is managed by considering the e-, d- and l-shifts as being of the same type. We denote this merged shift as M-shift and will refer to this transformation as transformation from the edlNR domain to the MNR domain.

In the MNR domain the requirement for staff cover during the corresponding shifts is summarised in Table 3. This in itself does not have any adverse effect on the computational complexity of the scheduling process. However, the important gain is that the reduction of the number of shifts from 5 to 3 makes the number of possible schedules in the MNR domain reduce to $28 * 3^{28} = 6 * 10^{14}$. This represents a reduction by a factor of 10^7 . There is a potential for additional domain transformation and the associated computational gain even though traditional scheduling methods are more efficient in this reduced space.

Table 3: Shift Types and the required numbers of nurses on specific shifts in the MNR domain

Shift type	Number of nurses on specific shifts						
	M	T	W	T	F	S	S
M	13	13	13	13	13	13	13
N	3	3	3	3	3	3	3
R	Notional shift that last minimum of 2 days						

We introduce a granulated data structure referred here as "pattern". Pattern represent a feasible sequence of shifts that has a specific cost associated with it. So, as for this NSP problem, there are 36 zero- cost patterns. Figure 2 provides examples of such zero-cost patterns and Figure 3 provides examples of non-zero-cost patterns.

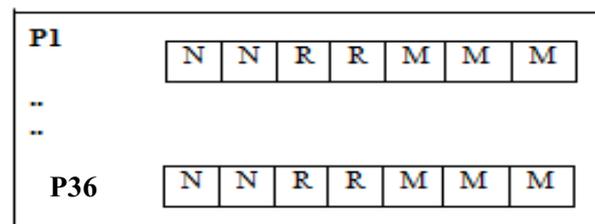


Figure2: No violation of Soft constraints (called as zero cost patterns)

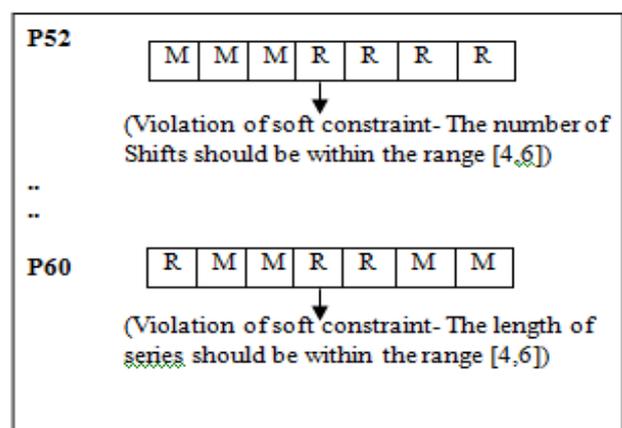


Figure3: Violation of Soft constraints with Cost 10 (called as non-zero cost patterns)

This solution of the scheduling problem in the domain of patterns needs to be converted back into the original edINR domain. This involves small computational effort primarily concerned with the consideration of the specific requirements with regard to the preference of e-, d- and l-shifts as summarised in Table 4.

Table 4: Penalties of violation for the conversion of MNR domain to edINR domain.

		Succeeding shifts			
		N	E	D	L
Preceding Shift	N	ok	n/f	n/f	n/f
	E	ok	ok	ok	ok
	D	nf	5	ok	5
	L	ok	ok	ok	ok

Early = E Day = D Late = L Night= N
Not Feasible = n/f

Our *domain transformation* approach can be summarised as a 3-stage process:

- I) conversion of the problem from the original edINR domain into a problem in the MNR domain;
- II) solution of the problem in the MNR domain
- III) conversion of the MNR solution into a solution in the original edINR domain

Integer Programming

In this paper, we apply Integer Programming an extension of Linear Programming that solves problem requiring integer solutions. We have solved the IP by implementing branch and bound. The basic concept underlying the branch-and-bound technique is to *divide and conquer*. To specify the above problem, the objectives are to minimize the values of individual variables. We formulate the entire problem associated with a 2-week scheduling period as the following IP model, which can be altered to adapt to any other problems with different constraints.

Let's call this binary pattern matrix: Pb .

This matrix is replicated for each nurse so the combined pattern matrix " Aeq ". The selection of patterns from the Aeq represents the schedule that satisfies the equality constraints such as the cover requirement. This can be expressed as:

$$Aeq' * x = c' \quad (2)$$

where x is the unknown binary vector, representing a solution to the scheduling problem and c is the staff cover requirement. The requirement that each nurse is assigned at most one pattern represents a constraint that can be written as

$$A' * x \leq b' \quad (3)$$

where A is a matrix with the number of columns corresponding to the number of nurses (say 28) and b is a vector of 1s corresponding to the number of nurses. The objective of the optimization of the scheduling might be defined as trying to satisfy the cover requirement with the minimum number of nurses. This is expressed simply as:

$$\text{Min } NP * x \quad (4)$$

where NP is a vector of 1s of size "number of nurses times the number of patterns". The mathematical model described above prepared a weekly schedule for wards up to 28 nurses at a government hospital at Kajang, Malaysia. The objective function is considered as a cost function, where cost is interpreted as penalty and penalty is defined based on the desirability of a nurse to work at a shift type on a day. Therefore our attempt is to minimize the penalty to the given constraints.

SIMULATION RESULT

For our approach, the IP part is solved by the latest GNU Octave's GLPK (4.45). For this simulation, an Intel Pentium 1.64GHz PC with 448MB RAM under Windows 7 was used. The results obtained by solving the Branch-and-Bound Integer Programming (BBIP) is presented in the following table. Practically, the materon may have to modify some shifts related to these violations, but these modifications are much easier than making a work table from scratch by hand. Solving within practical time will be dependent on the performance of IP solver. However, in the case of a problem with many shift types, the window width should generally be small.

Experiments on Scheduling Simulation

There are three experiments done according to different types of weeks. This simulation experiments done according to the domain transformation method. In order to verify the simulation model, hospital records are compared with simulation results. Most of the time, hospital did not manage to fulfil the demand that needs to be covered. However, by using our domain transformation approach, we manage to fulfill the demand; not just satisfying the hospital scheduling period of two weeks but also we proposed the four weeks and five weeks scheduling. Below are results on the final outputs generated. Table 5 to table 7 are results on the final outputs generated. They are presented in the Appendix. To facilitate this improvements, Table 8 shows the cost summary for the scheduling period of 2,4,and 5. It also shows the computational time required. This clearly shows that, our approach has the ability to introduce changes. Besides, it also can reduce computational complexity and therefore reduce the computational time.

Table 8: Table with time execution and cost summary

Weeks	Days	Cost	Time(s)
2	14	44	21
4	28	77	47
5	35	100	52

Numerical Result on Demand Simulation

Numerical experiments described in this section provide a representative sample of the simulation studies conducted to balance the degree of satisfaction of soft constraints vs. the decisions on employing additional nursing staff. We have varied the required cover on individual shifts to simulate the decision support functionality. Based on the original problem, we have changed few sample runs of different number of nurses Table 9 presents the results of the best set of nurses which satisfies the demand of the original problem with a very reasonable cost for a month. While Table 10 and 11 represents the alternative demands with the number of nurses and this is concluded in Graph 1, Graph 2 and Graph 3 which shows clearly the representation of the various cost.

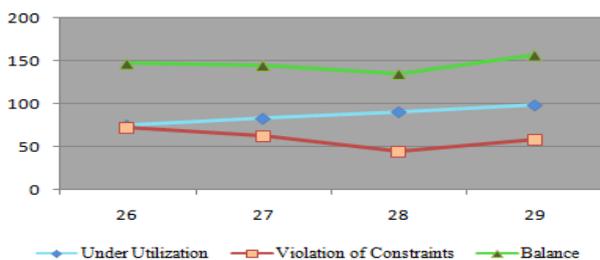
The results tables use the following notation:

- TN = Total number of nurses
- TC = Total number of contractual (in hours)
- TW = Total Number of hours worked(in hours)
- U(h/w) = Under Utilization of Nurses (hours/week)
- CSC = Cost of violating Soft constraint
- CU = Cost of under- utilization
- T(s) = Time (in seconds) to execute the software

Case 1:

Table 9: The balance of violation of soft constraints and the underutilisation of nurses for the “13131313131313” D-shift and the “3333333” N-shift cover (TW=667h)

TN	TC	U (h/w)	CSC	CU	Ctot	T(s)
26	1040	193	67	38.6	105.6	18
27	1080	233	58	46.6	104.6	20
28	1120	273	44	54.6	98.6	21
29	1160	313	52	62.6	114.6	35

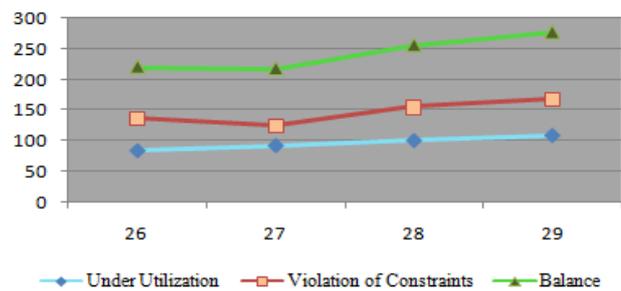


Graph 1: Balance between the constraint cost and the Under Utilization cost for the cover “13131313131313”

Case 2:

Table 10: The balance of violation of soft constraints and the underutilisation of nurses for the “12121212121212” D-shift and the “3333333” N-shift cover (TW=618h)

TN	TC	U (h/w)	CSC	CU	Ctot	T(s)
26	1040	422	136	84.4	220.4	45
27	1080	462	125	92.4	217.4	55
28	1120	502	155	100.4	255.4	57
29	1160	542	168	108.4	276.4	60

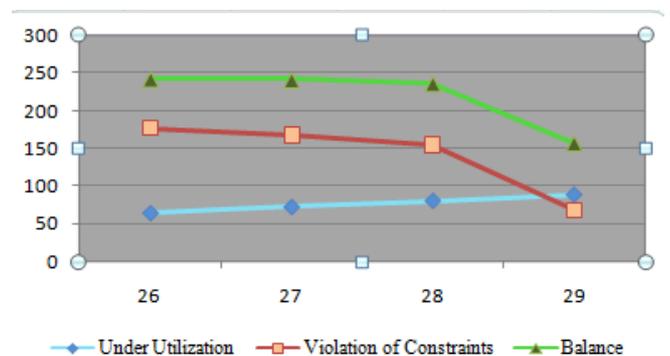


Graph 2: Balance between the constraint cost and the Under Utilization cost for the cover “12121212121212”

Case 3:

Table 12: The balance of violation of soft constraints and the underutilisation of nurses for the “14141414141414” D-shift and the “3333333” N-shift cover (TW=716h).

TN	TC	U (h/w)	CSC	CU	Ctot	T(s)
26	1040	324	177	64.8	241.8	64
27	1080	364	168	72.8	240.8	62
28	1120	404	155	80.8	235.8	56
29	1160	444	68	88.8	156.8	48



Graph 3: Balance between the constraint cost and the Under Utilization cost for the cover “14141414141414”

Discussion

This study illustrates the simulation modelling approach to nurse scheduling and the management decision support concerning the staffing levels. The scheduling simulation experiment provides the best schedules with different time horizons. Table 5,6, and 7 show snapshot of the schedules with the associated costs. Table 8 gives the summary of all the different time horizons and the associated costs. In reference to Table 8, we have suggested to the hospital management to have a 4 – 5 weeks scheduling as the cost is reduced with longer scheduling horizon. It is worth noting that manual scheduling by the hospital could not satisfy the required demand while our simulation result satisfies the demand and all other hard constraints.

Subsequently, the demand simulation experiments calculate how the constraints associated with the scheduling problem influence the cost-effectiveness of employing additional staff. We have shown this using a representative set of 3 different scenarios with different number of nursing staff considered in each scenario. The result indicates that for the original problem demand of “13131313131313” D-shift, the exact balance is 28 nurses, as indicated in the graph 1. With fewer than 28 nurses we can't satisfy the clinical cover requirement and having larger numbers of nurses implies unnecessarily higher employment cost. The balance for the alternative clinical cover requirements (demands) of “12121212121212” is 27 nurses; and for the “14141414141414” the required number of nurses is 29.

CONCLUSION

As a conclusion, nurses performance in a hospital can be managed and coordinated with the aid of nurse scheduling. Nurse scheduling helps departments in the hospital to organise the number of nurses working in a day either in day shift or night shift. In a nutshell, with proper scheduling method of the nurses, a high quality roster can be produced. From the research that has been carried out, it has been quite clear that preparing a nurse schedule requires a huge amount of assessment on various criteria's such as organizational rules, personal data, legal regulations and many more. No single software can be used since each hospital has its own set of requirements and constraints but the same method can achieve a good solution. For proper results, the models and algorithms involved in generating the schedule should have a strong yet flexible structure in order to adapt to various unexpected situations that occur in a hospital. Modeling and simulating both process of scheduling and staffing, provide decision makers with a specific system that observe the impact of both interlinked process flow types. In the case of the documentation aid, it is important to take both process flow types into account. Framework and guidance for the modeler are essential in order to develop the quality of the developed model. Therefore, the method

proposed is a valuable tool for the hospital modeler as it provides a raw model that can be adjusted to the requirements of the system under investigation. There are few main advantages that this simulated scheduling system offers such as it is cost effective, time efficient, repeatable and effective. This simulation model enables the visualization of the system over time. Besides, it is also versatile where model simulates real life and allows for a wide range of experiments with no impact on real objects. Domain transformation represents departure from a conventional one-shift-at-a-time scheduling approach. It offers an advantage of efficient and easily understandable solutions as well as offering deterministic reproducibility of the results. We note however that it does not guarantee the global optimum.

ACKNOWLEDGEMENTS

The authors would like to thank Kajang Hospital, Malaysia for providing us with an opportunity to participate in this project. Special thanks to the hospital matron Pn. Mashita binti Khalid for her guidance in providing the constraints. We would also like to thank Medical Research and Ethics Committee (MCEF) who gave approval to conduct this study, and Institute of Health Behavioural Research (IHBR) for allowing us to conduct this research in the respective hospital and include information on the hospital constraints in this paper. In particular, we would like to thank En.Nazrul and Cik Sharipazalira to get this paper approved for ECMS submission from the different committees. Last but not least, a special thanks to the Director General of Health to give us this opportunity to conduct this research in the Malaysia Hospital and for his permission to publish this article.

REFERENCES

- Abdennadher, S. and Schlenker, H. 1999. “Nurse scheduling using constraint logic programming”. IAAI99 Proceedings of the Sixteenth National Conference on Artificial intelligence and the Eleventh innovative Applications of Artificial intelligence Conference innovative Applications of Artificial intelligence, Orlando, Florida, United States, 1999, pp. 838-843.
- Bargiela, A., and Pedrycz, W. 2002. “Granular Computing – An Introduction”, Kluwer Academic Publishers. 2002. <http://dx.doi.org/10.1007/978-1-4615-1033-8>
- Bargiela, A. and Pedrycz, W. 2008. “Toward a theory of Granular Computing for human-centred information processing”, IEEE Trans. On Fuzzy Systems. 16(2): 320-330. <http://dx.doi.org/10.1109/TFUZZ.2007.905912>
- Baskaran Geetha, Bargiela Andrzej, Qu Rong. 2009. “Hierarchical method for nurse rostering based on granular pre-processing of constraints”, Proc. 23rd European Conference on Modelling and Simulation, ECMS, Madrid, Spain, pp.855-861, June 2009, (doi: 10.7148/2009-0855-0861).
- Baskaran, G., Bargiela, A., Qu, R. 2012. “From Simplified to Detailed Solutions to the Nurse Scheduling Problem”, 25th European Conference on Operational Research,

EURO, Session TC-14, p.134-135, Vilnius, Latvia, July 2012.

Baskaran Geetha, Bargiela Andrzej, Qu Rong. 2013. "A Study of Cost Effective Scheduling of Nurses Based on the Domain Transformation Method", Proc. 27th European Conference on Modelling and Simulation, ECMS, Alesund, Norway, pp.309-314, May 2013, (doi: 10.7148/2013-0309).

Bouarab, H., Champalle, S., Dagenais, M., Lahrichi, N., Legrain, A., and Taobane, M., (2010). Nurse Sceduling: From Theoretical Modeling to Practical Resolution. (Resource Optimization in Healthcare).Fall Brucker, P., Glass, C. A., and R. A. Knight. 2010. "The nurse rostering problem: A critical appraisal of the problem structure". European Journal of Operational Research 202 (2): 379 – 389.

Henderson, V.A and Harmer, B. 1939. "The principles and Practice of Nursing". (4th ed). New York: Macmillan.

Qu,R., Burke,E., and Post,G. "A decomposition, construction and post-processing approach for a specific nurse rostering problem," in Proceeding of the 2nd Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA'05), New York, USA, 2005 pp. 397-406.

Sitompul, D. and Randhawa,S. "Nurse scheduling models: a state-of-the-art review,"Journal of the Society of Health Systems, vol. 2, pp. 62-72, 1990.

AUTHOR BIOGRAPHIES



GEETHA BASKARAN was born in Melaka, Malaysia. She is an Assistant Professor at the University of Nottingham, Malaysia Campus. She is a member of the Automated Scheduling

and Planning research group in the School of Computer Science at the University of Nottingham. She is currently pursuing her PhD studies here focussing on Nurse Scheduling. Her main research area include nurse scheduling, domain transformation, information granulation, heuristic, IP,ILP, matrix exploration.



ANDRZEJ BARGIELA is Professor in the School of Computer Science at the University of Nottingham. He served as President of the European Council for Modelling and Simulation

(ECMS) during 2002-2006 and 2010-2012. He is Associate Editor of the IEEE Transactions on Systems Man and Cybernetics and Associate Editor of the Information Sciences. His research involves investigation into Granular Computing, human-centred information processing as a methodological approach to solving large-scale data mining and system complexity problems.



DR RONG QU is a Lecturer in the School of Computer Science at the University of Nottingham. She gained

her PhD in Computer Science from the University of Nottingham in 2002. Her main research areas include meta-heuristics, constraint programming, IP/ILP, case

based reasoning methodologies and knowledge discovery techniques on scheduling, especially educational timetabling, healthcare personnel scheduling, network routing problems and graph colouring. In total she has more than 30 papers published or to appear at international journals and peer-reviewed international conferences. Dr Qu is also a guest editor for special issues at the journal of Memetic Computing and the Journal of Scheduling, and the program chair of several workshops and an IEEE symposium.

APPENDIX

Table 5: Two week in **edINR** schedule with total cost 44

Computed schedule for 2 week(s)									
Days->	MTWTFSS	viol	cost	MTWTFSS	viol	cost			
Nurse 01	(40,)	:LLNNRRR	0	0	REEEERR	0	0		
Nurse 02	(40,)	:NNRRELE	0	0	EEEEERE	0	0		
Nurse 03	(40,)	:RRRLNNN	0	0	RRREEER	0	1		
Nurse 04	(40,0)	:DDDDRRR	0	0	RRRDDDD	0	0		
Nurse 05	(40,)	:RREEEER	0	0	LENNRRR	0	0		
Nurse 06	(40,)	:EEERRRL	0	0	EEEEERR	0	0		
Nurse 07	(40,)	:EEERRLE	0	0	NNRRREE	0	0		
Nurse 08	(40,)	:EERRRLE	0	0	NNRRREE	0	0		
Nurse 09	(40,)	:RRRELEE	0	0	EERRREE	0	0		
Nurse 10	(40,)	:LLNNRRR	0	0	RDEEERR	0	5		
Nurse 11	(40,)	:NNRREEE	0	0	ERRRELE	0	0		
Nurse 12	(40,)	:RREEEER	0	0	RRDEELR	0	5		
Nurse 13	(40,)	:RREEEER	0	0	LLNNRRR	0	0		
Nurse 14	(40,)	:EEERRRD	0	0	ELLRRRL	0	5		
Nurse 15	(40,)	:RELEDRL	0	0	LLNNRRR	0	0		
Nurse 16	(40,)	:LLNNRRR	0	0	RLLLLR	0	0		
Nurse 17	(40,)	:NNRRELE	0	0	DRRLLL	0	0		
Nurse 18	(40,)	:ERRRNNN	0	0	RRLLLR	0	10		
Nurse 19	(40, S)	:EEEEERR	0	0	RRRLN	0	1		
Nurse 20	(40, S)	:RRLLEER	0	0	ELEEER	0	0		
Nurse 21	(40, S)	:LLLRRRE	0	0	LELRRE	0	0		
Nurse 22	(40, S)	:LLRRLL	0	0	NNRRLLL	0	0		
Nurse 23	(40, S)	:LLLRRRL	0	0	LLRRLL	0	0		
Nurse 24	(40, S)	:RRLLLR	0	0	RRRLN	0	1		
Nurse 25	(40, S)	:RRRRRR	0	10	RRRRRR	0	10		
Nurse 26	(40, S)	:RRRLN	0	0	RRLLLR	0	1		
Nurse 27	(40, S)	:RRLLDL	0	0	LRNN	0	5		
Nurse 28	(40, S)	:RRLLLR	0	0	RRLLLR	0	0		

Verifying total nurses available each day:										
Total E:	6666666		6666666							
Total D:	1111111		1111111							
Total L:	6666666		6666666							
Total N:	3333333		3333333							

Table 6: Suggested schedule 4 week in edINR schedule with total cost 77

Computed schedule for 4 week(s)

Days->	MTWTFSS	viol	cost	MTWTFSS	viol	cost	MTWTFSS	viol	cost	MTWTFSS	viol	cost
Nurse 01(40,)	:LLNNRRR	0	0	REEEERR	0	0	RRRLNNN	0	0	RREEEER	0	1
Nurse 02(40,)	:NNRRELE	0	0	EEERRLL	0	0	NNRRREL	0	0	EEEEERR	0	0
Nurse 03(40,)	:RRRLNNN	0	0	RREEEER	0	0	RREEEER	0	0	EEEEERR	0	1
Nurse 04(40,0)	:DDDDRRR	0	0	RRRDDDD	0	0	RRDDDDR	0	0	RDDDDRR	0	0
Nurse 05(40,)	:RREEEER	0	0	LENNRRR	0	0	EEEEERR	0	0	RRRLNNN	0	1
Nurse 06(40,)	:EEERRRL	0	0	EEEEERR	0	0	LENNRRR	0	0	RREEEER	0	0
Nurse 07(40,)	:EERRRLE	0	0	NNRRREL	0	0	EEEEEEE	0	0	ERRRNNN	0	0
Nurse 08(40,)	:ERRRLLR	0	0	NNRRREL	0	0	EEEEERR	0	0	RREEEER	0	0
Nurse 09(40,)	:RRRELEE	0	0	EERRRLL	0	0	NNRRREE	0	0	EERRREE	0	0
Nurse 10(40,)	:LLNNRRR	0	0	RDEEERR	0	0	RRLNNN	0	0	RRRLEEE	0	5
Nurse 11(40,)	:NNRREEE	0	0	ERRRLEE	0	0	ERRRLEE	0	0	ERRRLEE	0	0
Nurse 12(40,)	:RREEEER	0	0	RRDEEER	0	0	LENNRRR	0	0	RRELLDR	0	5
Nurse 13(40,)	:RREEEER	0	0	LLNNRRR	0	0	RRRLLE	0	0	NNRRRLE	0	0
Nurse 14(40,)	:EEERRRD	0	0	ELLRRRE	0	0	EDERRRE	0	0	DELRRRD	0	5
Nurse 15(40,)	:RELEDR	0	0	LLNNRRR	0	0	RLEEERR	0	0	LLNNRRR	0	0
Nurse 16(40,)	:LLNNRRR	0	0	RLLLLRR	0	0	RLEEERR	0	0	LLNNRRR	0	0
Nurse 17(40,)	:NNRRLEL	0	0	DRRRLLE	0	0	DRRRNNN	0	0	RRLLLRL	0	10
Nurse 18(40,)	:ERRRNNN	0	0	RRLLLRL	0	10	NNRRLLR	0	10	RLLRREL	0	11
Nurse 19(40, S)	:EEEEERR	0	0	RRRLNNN	0	0	RRREEEE	0	0	LRRREEL	0	1
Nurse 20(40, S)	:RRLLEER	0	0	ELEEEER	0	0	LLNNRRR	0	0	RRELELR	0	0
Nurse 21(40, S)	:LLLRRRE	0	0	LELRRRE	0	0	EEERRLL	0	0	EEERRLL	0	0
Nurse 22(40, S)	:LLLRRLL	0	0	NNRRLLE	0	0	LRRRLLL	0	0	NNRRLLL	0	0
Nurse 23(40, S)	:LLLRRRL	0	0	LLLRRRL	0	0	LLLRRRL	0	0	LLNNRRR	0	0
Nurse 24(40, S)	:RRLLLLL	0	0	RRRLNNN	0	0	RRRLLLL	0	0	RLLLLRR	0	1
Nurse 25(40, S)	:RRRRRRR	0	10	RRRRRRR	0	10	RRRRRRR	0	10	RRRRRRR	0	10
Nurse 26(40, S)	:RRRLNNN	0	0	RRRLLLL	0	0	LLRRLLL	0	0	NNRRLLE	0	11
Nurse 27(40, S)	:RRRLLDL	0	0	LRRRNNN	0	0	RRLLRD	0	10	LRRRNNN	0	15
Nurse 28(40, S)	:RRLLLLR	0	0	RRLLLLR	0	0	RRLLLLR	0	0	LLLRLLL	0	0

Verifying total nurses available each day:

Total E:	6666666		6666666		6666666		6666666
Total D:	1111111		1111111		1111111		1111111
Total L:	6666666		6666666		6666666		6666666
Total N:	3333333		3333333		3333333		3333333

Table 7: Suggested schedule 5 week in edINR schedule with total cost 100

Computed schedule for 5 week(s)

Days->	MTWTFSS	viol	cost	MTWTFSS	viol	cost	MTWTFSS	viol	cost	MTWTFSS	viol	cost
Nurse 01(40,)	:LLNNRRR	0	0	REEEERR	0	0	RRRLNNN	0	0	RREEEER	0	1
Nurse 02(40,)	:NNRRELE	0	0	EEERRLL	0	0	NNRRREL	0	0	EEEEERR	0	0
Nurse 03(40,)	:RRRLNNN	0	0	RREEEER	0	0	RREEEER	0	0	EEEEERR	0	1
Nurse 04(40,0)	:DDDDRRR	0	0	RRRDDDD	0	0	RRDDDDR	0	0	RDDDDRR	0	0
Nurse 05(40,)	:RREEEER	0	0	LENNRRR	0	0	EEEEERR	0	0	RRRLNNN	0	1
Nurse 06(40,)	:EEERRRL	0	0	EEEEERR	0	0	LENNRRR	0	0	RREEEER	0	1
Nurse 07(40,)	:EERRRLE	0	0	NNRRREL	0	0	EEEEEEE	0	0	ERRRNNN	0	10
Nurse 08(40,)	:ERRRLLR	0	0	NNRRREL	0	0	EEEEERR	0	0	RRRLLE	0	0
Nurse 09(40,)	:RRRELEE	0	0	EERRRLL	0	0	NNRRREE	0	0	EERRRLE	0	0
Nurse 10(40,)	:LLNNRRR	0	0	RDEEERR	0	0	RRLNNN	0	0	RRRLEEE	0	5
Nurse 11(40,)	:NNRREEE	0	0	ERRRLEE	0	0	ERRRLEE	0	0	ERRRLEE	0	0
Nurse 12(40,)	:RREEEER	0	0	RRDEEER	0	0	LENNRRR	0	0	RREEEER	0	5
Nurse 13(40,)	:RREEEER	0	0	LLNNRRR	0	0	RRRLLE	0	0	EELERR	0	0
Nurse 14(40,)	:EEERRRD	0	0	ELLRRRE	0	0	EDERRRE	0	0	EELRRRD	0	5
Nurse 15(40,)	:RELEDR	0	0	LLNNRRR	0	0	RLEEERR	0	0	LLNNRRR	0	5
Nurse 16(40,)	:LLNNRRR	0	0	RLLLLRR	0	0	RLEEERR	0	0	LLNNRRR	0	0
Nurse 17(40,)	:NNRRLEL	0	0	DRRRLLE	0	0	DRRRNNN	0	0	RRLLLRL	0	10
Nurse 18(40,)	:ERRRNNN	0	0	RRLLLRL	0	10	NNRRLLR	0	10	RLLRREL	0	10
Nurse 19(40, S)	:EEEEERR	0	0	RRRLNNN	0	0	RRREEEE	0	0	LRRREEL	0	1
Nurse 20(40, S)	:RRLLEER	0	0	ELEEEER	0	0	LLNNRRR	0	0	RRELELR	0	0
Nurse 21(40, S)	:LLLRRRE	0	0	LELRRRE	0	0	EEERRLL	0	0	EEERRLL	0	0
Nurse 22(40, S)	:LLLRRLL	0	0	NNRRLLE	0	0	LRRRLLL	0	0	NNRRLDL	0	5
Nurse 23(40, S)	:LLLRRRL	0	0	LLLRRRL	0	0	LLLRRRL	0	0	LLNNRRR	0	0
Nurse 24(40, S)	:RRLLLLL	0	0	RRRLNNN	0	0	RRRLLLL	0	0	RLLLLRR	0	1
Nurse 25(40, S)	:RRRRRRR	0	10	RRRRRRR	0	10	RRRRRRR	0	10	RRRRRRR	0	10
Nurse 26(40, S)	:RRRLNNN	0	0	RRRLLLL	0	0	LLRRLLL	0	0	EEERRRL	0	11
Nurse 27(40, S)	:RRRLLDL	0	0	LRRRNNN	0	0	RRLLRD	0	10	LRRRNNN	0	15
Nurse 28(40, S)	:RRLLLLR	0	0	RRLLLLR	0	0	RRLLLLR	0	0	RRRLNNN	0	1

Verifying total nurses available each day:

Total E:	6666666		6666666		6666666		6666666		6666666
Total D:	1111111		1111111		1111111		1111111		1111111
Total L:	6666666		6666666		6666666		6666666		6666666
Total N:	3333333		3333333		3333333		3333333		3333333

EMERGENCY DEPARTMENT: A GENERAL ADAPTABLE SIMULATION MODEL IMPLEMENTED IN ARENA

Arturo Liguori

Epidemiology and Community Medicine Unit,
Department of Paediatrics - University of Padova
Via Pietro Donà 11, 35129 Padova, Italy
e-mail: epi@pediatria.unipd.it
phone: +39-049-8215700, fax: +39-049-8215700

Giorgio Romanin-Jacur

Department of Management and Engineering
University of Padova
Stradella San Nicola 3, 36100 Vicenza, Italy
e-mail: romjac@dei.unipd.it
phone: +39-335-6072747, fax: +39-0444-998888

Abstract- The Emergency Department of a hospital is devoted to provide first aid to outpatients. A correct organization and resource dimensioning is very important both in the planning and in the management phase and may be usefully supported by a simulation model to be applied by administrators and operators. As an emergency department is a very complex framework, a model which simulates it requires a large amount of time and an expert software programmer to be built and implemented. In this paper a generalized flexible model has been built up, able to reproduce all common structural and functional characteristics of every actual emergency room. This simulation model can be easily adapted to almost all emergency departments only by defining its functional parameters without altering its structure; it is written in language SIMAN by tool Arena, widely diffused, and permits an easy readability also by non expert users.

Keywords- Emergency department, parametric model, discrete simulation.

1 INTRODUCTION

The emergency department is a hospital department devoted to provide first aid to outpatients who suffer from an injury or an illness requiring urgent care.

Emergency department service is characterized by high variability of patient arrivals, depending on time (different intervals of the day, week, year), according to extreme randomness; moreover the required assistance type may vary according to the patient characteristics and to the suffered injury or illness. On the patient side, a response which shall be quick enough (sometimes immediate) with respect to patient state severity, is expected, also in case of congestion. As a consequence of all above needs, an emergency department shall be correctly designed and managed for what concerns structures, technological resources and human resources, in order to supply a high quality service at minimum cost.

Once given the presence of randomness in patient arrival, in patient management (due to priorities and possible preemption) and in services' duration, an analytical model (see [16]) is not suitable, as it is able to provide only mean behaviour results, while a simulation model, able to describe in detail system behaviour, and to give results related to

extreme conditions, appears to be the most convenient, as shown by a wide literature. Many papers using simulation are dedicated to a generic emergency department to solve specific management problem (see [4], [11], [9], [26], [18], [31], [33], [17], [20], [32], [12]) or to study the effects of triage organization (see [7]); paper [25] studies the behaviour of a generic emergency department in the case of a maxi-emergency; the above papers are addressed to analyze generic departments to solve specific problems. Other papers solve problems of the same kind but applied to specific departments (see [23], [30], [3], [10]); in particular [10] implements the model in simulation tool Arena, which is particularly friendly and understandable also by non experts. In some papers the simulation model describes also (or only) other services connected with the emergency departments, like ambulances and patient transportation, either in a generic situation (see [19], [27], [5], [1], [21], [3]) or in a specific one ([28], [15], [6]); paper [21] studies interaction among emergency department and other services or departments inside and outside the hospital. Paper [13] suggest a generalized model of emergency department without discussing its implementation. A side consideration is due to papers dedicated to triage ([2], [8], [14], [22], [24]).

In the present paper a generalized model of emergency department, flexible and adaptable to the majority of existing departments, is presented; its adaptability permits to model and simulate both departments managing ambulances with doctors and/or nurses and departments independent from patients' transportation systems; patients' arrivals probability distributions, both in normal and in exceptional conditions, are characterized by adjustable parameters; structural and human resources management, turns of duty and priority rules can be ruled so to reproduce the behaviour of the studied service (see figure 1 for general flow chart model). The model implementation is developed by tool Arena, widely used in the world and very easy to be understood by non expert users, like medical personnel, with whom the model shall be adapted and employed to adjust department dimensions and to improve department effectiveness and efficiency.

For what concerns the paper organization, the department operation and the generalized model is discussed in Section 2 while its implementation in Arena is explained in Section 3; results and possible suggestions are reported in Section 4. Conclusion follows in Section 5.

2. EMERGENCY DEPARTMENT OPERATION AND GENERALIZED MODEL

If we consider the emergency department from the patients' point of view, we may see that their movements may be outlined as follows:

- arrival;
- triage (after possible wait in queue);
- a sequence of medical examinations (with possible intervention and possible resuscitation), and exams or specialized consulting, alternatively followed by: sending home, hospital admission, short term admission; often examinations require wait in queue; sometimes the sequence may be interrupted by patient's death in severe cases;
- a final examination, followed either by sending home or by hospital admission, at the end of short term admission, if occurred.

The presented model may be considered to be absolutely general and it summarizes all application experience of last years (11 specific departments in Italy and abroad, in normal and exceptional situations), as the above actions' sequence (triage, different examinations and exams) is widely adopted in all emergency departments. The setting of all parameters ruling all phases' durations however permits to activate or deactivate single phases in correspondence of specific applications.

The general sequence of arrival, triage and examinations is represented in figure 1. The required resources are:

- structural resources: major and minor treatment rooms, waiting rooms, short term admission rooms;
- technological resources: specific instruments, which are generally strongly connected with the structural resources, and therefore are not considered separately;
- human resources: medical doctors, general and specialist nurses, auxiliary operators, working according to well defined rules and turns of duty.

Resources employed to perform single tasks are chosen from available resource sets, according to predefined rules stating priority, compatibility and suitability.

Starting from this very general model with low detail, every function will be explained in detail.

2.1 Patients arrivals

We may classify arrivals as normal arrivals, special arrivals and pseudo-arrivals.

Normal arrivals are related with either traumatic accidents or urgent medical need, due for instance to infarct, stroke or severe health worsening; their happening is absolutely random and independent, even if a strong dependency on time (day's intervals and weekdays) and on patient type (as will be specified later) is present in the mean. Normal patients may arrive by their own means (on foot, by car) or by an ambulance, which may be provided either by the same department (equipped or not with a doctor or a nurse) or by a separate rescue organization.

Special arrivals are related with exceptional situations which cause abnormal increase in patient arrivals for a limited time period, and may be divided in two groups.

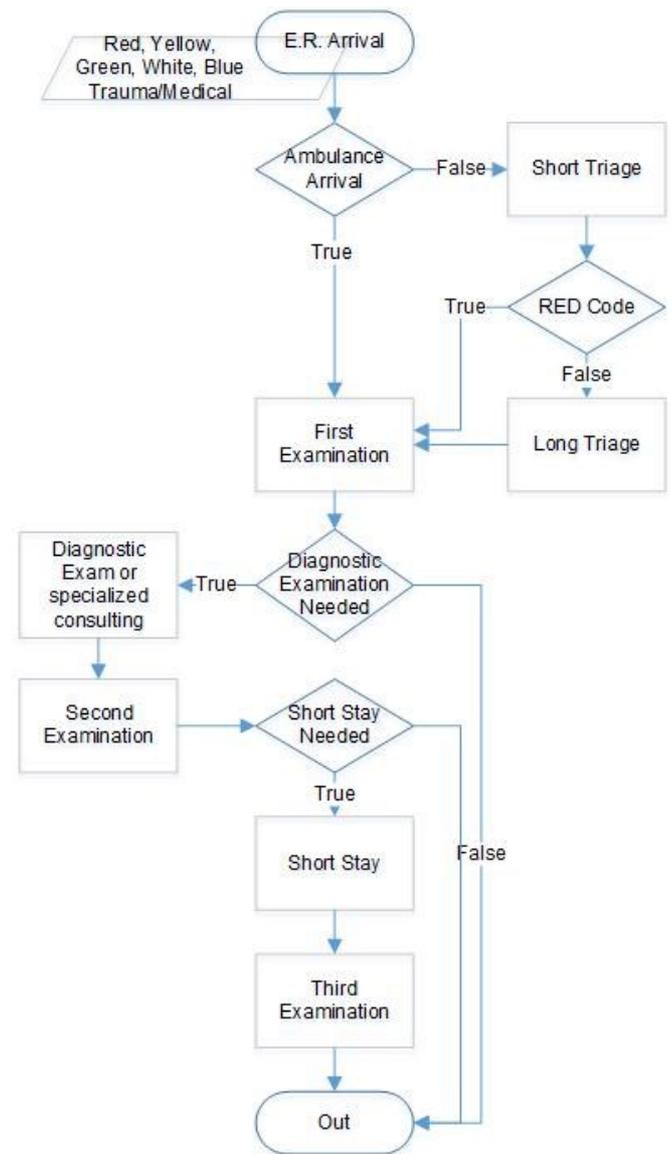


Figure 1: Generalized flowchart diagram

In the first group we classify patients generated either by extraordinary weather (high temperature, dryness, high humidity, low temperature) or by epidemics (typically flu) who generally are either old or already weakened by other illness; their arrivals are random but depend on time according to the specific cause. In the second group we classify patients generated by a maxi-emergency, which happens for instance because of severe road accidents involving many vehicles, fires, building collapses, earthquakes, spread of toxic gas, etc.; such patients may be in a large number (from tens to hundreds) and arrive in a very short interval (from an hour to six hours), and therefore they evidence a brutal, even if temporary, insufficiency of

immediately available rescue with respect to assistance requests, to which the department shall react with a rapid pre-ordered outstanding effort; the maxi-emergency occurrence is rare and random, the patients' generation has been classified, according to the cause, as minimal (48 patients of various types in an hour), strong flow-short term (100 patients of various types in an hour) or long term (272 patients of various types in 6 hours) .

Pseudo-arrivals are assistance requests by people admitted in the short term admission rooms; such rooms are dedicated to short stays and are managed by the same emergency department doctors' and nurses' team; the assistance request are random independent but their mean is proportional to the number of currently admitted patients.

The mean arrival time depends on the type of patient and is generally obtained by a suitable statistical investigation.

2.2 Triage

All arriving patients are immediately selected by the same unique filter (triage), which is ruled by specialized nurses. The selection is based on the presented life parameters, predefined and universal, breathing, heartbeat and consciousness, which determine the immediate life danger.

The selection assigns a colour code, according to immediate life danger, and related urgency:

- red code, aid shall be immediate, as any wait may cause death;
- yellow code, aid shall not be delayed more than some minutes, wait may increase severity, a frequent check is required if assistance cannot be immediate;
- green code, aid may be delayed, a limited wait (tenth of minutes to two hours) is permitted without danger; anyway assistance cannot be supplied out of the hospital;
- white code, no urgency and no danger, any wait cannot cause damage; assistance should be supplied out of the emergency department (general practitioner or hospital ambulatory).

Beside the upper colour codes, there is the blue code, which is assigned to patients who received assistance by the emergency department in previous days and return for checks (for instance for stitches on a wound).

The triage operation is executed in two phases:

- a first look, which requires a very short time, e.g. one minute, only to reveal a possible red code, and in such a case assistance is immediate; if all rooms, doctors and nurses are busy with lower code patients, pre-emption takes place, i.e., other actions are interrupted to assist the red code patient; the first look is executed by the ambulance personnel, if the patient arrives by ambulance;
- a second look, where the patient (of yellow, green or white code) gets measurements (temperature, blood pressure, electrocardiogram if necessary, ...), first simple assistance (e.g., bandage) when necessary, and the nurse collects patient's personal data and information about his/her injury or illness. After second look triage the patient is put in a waiting queue and remains in a waiting room under the supervision of triage nurse(s); the complete assistance will be

obtained according to colour priority (yellow, green, blue, white) as soon as a suitable room and a suitable doctor-nurse team is idle. During the waiting period a health worsening is possible: in such a case the triage nurse(s) shall upgrade the code and proceed accordingly.

Blue code patients do not get triage and are directly put in a waiting queue with priority on white code patients.

2.3 Examination sequence

Patients of red code are assisted by following a fixed examination sequence:

- possible pick up by the department ambulance, equipped with a department doctor or a nurse according to department rules, if the department manages ambulances;
- first look triage, possibly in ambulance;
- first examination, obtained immediately, also by pre-emption with respect to lower code patients being assisted when no assisting personnel is idle; by such examination possible resuscitation, stabilization and assistance are provided; a major treatment room and a suitable doctor-nurse(s) group is required, according to the department rules; major assisting rooms may be either generic for red codes or specifically devoted to traumatic or medical assistance, according to department logistics; at the end the assisting doctor(s) alternatively decide(s) about: a) return home, b) hospital admission, c) short term room admission, d) further exams or consultations; patient's death may interrupt the examination;
- possible exams or consultations to investigate the patient's clinical picture; exams may be effected either inside the department (e.g. urine or blood tests) or in a laboratory; consultations are generally obtained by calling a hospital specialist (e.g. a cardiologist or a neurologist);
- second examination after exams' or consultation results, after which the assisting doctor(s) alternatively decide(s) about: a) return home, b) hospital admission, c) short term room admission; patient's death may interrupt the examination;
- third examination after short stay; it is to be specified that admission in short term room is chosen whenever a quick evolution of the clinical state is expected, so to decide for either home or hospital; short stay is limited to a maximum of 36-48 hours, but is generally shorter.

Patients of yellow code are assisted by following a different examination sequence:

- possible pick up by the department ambulance, equipped with a department doctor or a nurse according to department rules, if the department manages ambulances;
- first look triage, possibly in ambulance;
- second look triage and wait in the waiting room;
- first examination, obtained with priority with respect to lower code patients; by such examination stabilization and assistance are provided; a major treatment room and a suitable doctor-nurse(s) group is required, according to the department rules; major assisting rooms may be either generic for yellow (sometimes for both red and yellow) codes or specifically devoted to traumatic or medical

assistance, according to department logistics; at the end the assisting doctor(s) alternatively decide(s) about: a) return home, b) hospital admission, c) short term room admission, d) further exams or consultations; patient's death may interrupt the examination;

- possible exams or consultations to investigate the patient's clinical picture; exams may be effected either inside the department (e.g. urine or blood tests) or in a laboratory; consultations are generally obtained by calling a hospital specialist (e.g. a cardiologist or a neurologist);

- second examination after exams' or consultation results, after which the assisting doctor(s) alternatively decide(s) about: a) return home, b) hospital admission, c) short term room admission; patient's death may interrupt the examination;

- third examination after short stay.

Patients of green and white codes are assisted by a sequence which differs from the one of yellow codes in the choice of the room, which may be either a major treatment or a minor treatment one, and sometimes may be in common with rooms devoted to yellow codes; because of lower urgency, consultations are obtained in specialized departments of the hospital by moving the patient instead of calling a specialist; an obvious difference lays in the assistance priority, so that the queue time may be longer for green codes and much longer for white codes; finally, patient's death is very rare.

Patients of blue code get only one short visit.

During assistance sequence a change of patients severity (change of colour) may occur; in such a case the operations related to the new colour are applied.

The duration of various examinations depend on: 1) patient colour, 2) traumatic or medical assistance required, 3) composition of assisting team, 4) examination order (first, second, third); the most suitable statistical time distribution is a gamma one.

2.4 Employed resources

Emergency department organization always subdivides rooms into colour areas: a red area, devoted to red codes; a yellow area for yellow codes; a green area for green codes and a blue-white area for blue and white codes. In case of reduced department operations amount, some areas may be grouped together, for instance red area, green-yellow area and blue white area; for large operations amount, each area may be further subdivided into rooms devoted to traumatic and medical assistance. Room arrangement may change during the day, for instance some rooms may be closed and other rooms grouped together during the night. New rooms may be opened in case of a special patients arrival or a maxi-emergency, according to a pre-ordered plan.

For what concerns doctors' and nurses' staff, every person has his/her turns of duty, during which he/she is devoted to a specified area; rules are pre-defined for possible changes of area, whenever necessary (for instance from yellow to red area). Auxiliary people are dedicated to patients movements. In case of special arrivals or maxi-emergency, "available" doctors and nurses are called (according to a pre-ordered

procedure) and arrive in a short time (typically, half an hour) to increase the number of present ones.

3 SIMULATION MODEL IMPLEMENTATION IN ARENA

For the above model implementation Arena simulation tool was chosen, by considering that it is widely used in the world, and is easily read also by non users, due to animation and clarity of graphical representation.

The developed model uses about 250 blocks, of which about 40 creation blocks, about 100 process blocks, about 100 assign blocks and about 60 decide blocks; the model employs about 5 sub-models; obviously the small amount of different blocks increases readability.

The model may be split into three parts.

In the first part patient arrivals and pseudo-arrivals are generated.

In order to obtain the maximum model flexibility we chose to implement a different input for every possible patient type, and define for the corresponding arrivals a specific (either random or deterministic) scheduling. If a patient type is not to be considered, then the corresponding scheduler is set to zero.

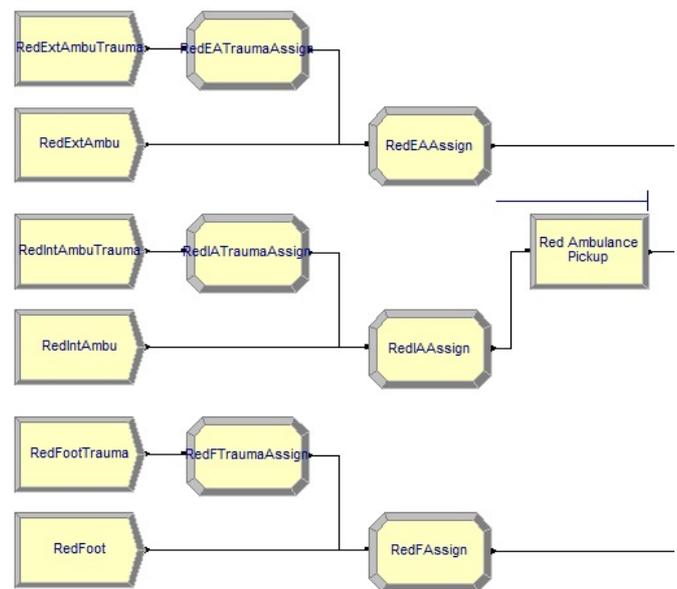


Figure 2: Red Code arrivals full schema

A set of creation blocks generate all types of normal patient arrivals, more precisely red, yellow, green and white codes, each one in all possible versions, i.e., arrival by department ambulance, by external ambulance and by foot, i.e., individual transportation means; blue codes arrive only by foot; all generations are modulated by a different scheduler, which permits to obtain random independent arrivals whose inter-arrival time may vary during the days of a week. To every patient a first vector, reporting colour, priority and next step is assigned. Patients arriving by department ambulance may employ a doctor or a nurse for the rescue trip, according

to a “preferred order” which is defined by the department rules. Obviously not all patient types shall be activated: for instance if the considered department does not manage ambulances, then all codes arriving by inner ambulance are not created by setting the related schedulers to zero.

Arrivals are divided by colour classification **Code** (Red, Yellow, Green, White, Blue), **type** (traumatic or medical patients) and E.R. **arrival mode** (internal or external ambulance, autonomous) (see figure 2 for the detail of Red Code Arrivals); such a complex distribution of arrivals, although it is not applicable to E.R. with an easier management of patients and to those that do not have their own ambulances, allows better and more accurate resource management.

The first Create Block rules the arrival of the patients associated with a specific schedule (see figure 3) that represents the number of arrivals expected by color, type and arrival way.

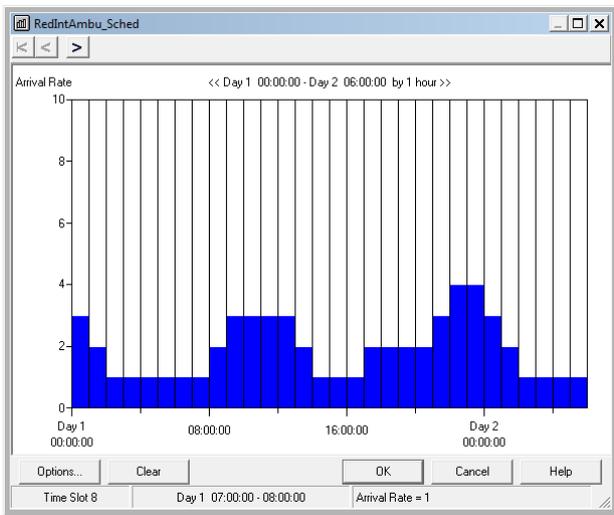


Figure 3: arrival schedule representation

The Assign Block assigns to each entity some attributes (as may be seen in the following list) which will address entities in the right way.

Attribute	PatientCode	"Red"
Attribute	PatientPriority	RedPriority
Attribute	Ambulance	1
Attribute	IntAmbu	1
Attribute	PatientPathState	"3FirstVisit"
Entity Picture	Picture.Red Ball	1
Attribute	AccompanimentNeeded	%RedAccompaniment

“Ambulance Pickup” Process (see figure 2) Seizes and Releases the resource that is needed in case an Internal Ambulance is sent to pick-up the patient.

A different creation block generates special arrivals (see figure 4), scheduled on a year time interval; every special arrival in its whole both generates single patients, who are addressed to increase the respective normal patients, and causes a suitable call for new available doctors nurses and rooms, whose amount is proportional to the number of special patients arrived.

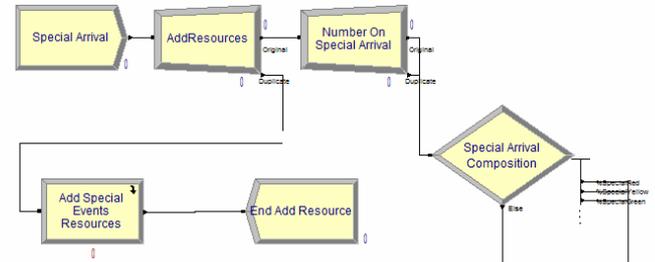


Figure 4: special arrivals block schema

Arrivals are scheduled on an annual basis and then a separate module and a following “N-Way by chance” decision module split the arrivals simulating (in a parametric way) the arrival of a certain number of patients with different color code distribution.

“Add Special Event Resources” Submodel is needed to introduce in the simulation additional resources (Doctors, Nurses, Rooms/Beds) for a limited period or until there is at least one special entity in the model. This settings are parametrical decided by a group of variables (see figure 5).

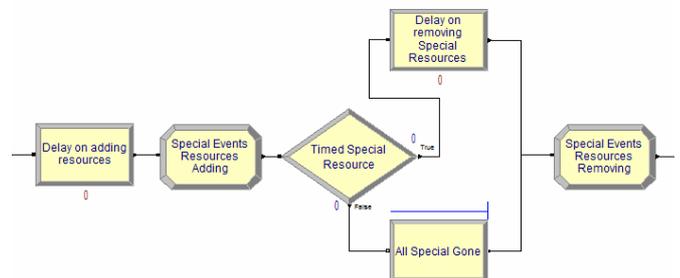


Figure 5: special resources adding submodel detail

Finally pseudo arrivals are generated by a creation block where the mean is modulated by a decide block, in order to be proportional to the number of occupied beds in the short term room (see figure 6).

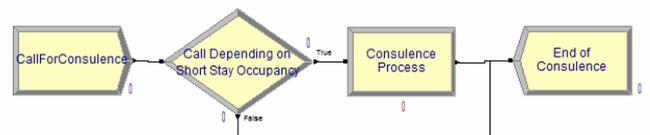


Figure 6: pseudo arrivals model

Pseudo Arrivals are scheduled on a day period, using the parameter $ResUtil(ShortStayBed)*100$ it is possible to

discard a number of calls according to the number of short stay bed occupancy.

In the second part triage is reproduced.

Red code patients arrived by ambulance skip the triage, while red codes arrived by individual means limit their us to first look triage. All other codes perform first look and second look triage, but the examination may be interrupted by the arrival of a red code; such an interruption is obtained by building a sub-model where the examination time is fractioned in a number of (parametrically predefined) time units and the examination can be broken in correspondence of a unit end, whenever required, and reset just after the interrupting patient has been served (triage simulation schema is shown in figure 7).

Apart from decision boxes using entity color code attribute to determinate the right path for each entity and assign boxes we have two process:

- short triage, a simple Seize Delay and Release process used to reserve the right resource (triage nurse or any available Nurse choosen with priority from a set)
- a long triage subprocess whose detail is quite similar to the examination process

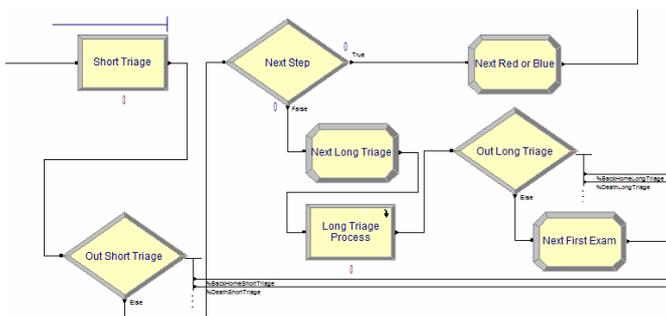


Figure 7: Triage Model

In the third part the assistance sequence is reproduced.

The first and possible second examination seen in the previous paragraph (as shown in figure 8) are simulated by sub-models similar to the one reproducing second look triage, in order to permit pre-emption whenever required; examinations' duration is dependent on patient's characteristics, colour, examination order, traumatic or medical, and ruled by the related parameters. Possible inner/outer exams and consultations, with different parameters according to the patient's characteristics, are represented; required patients' transportations are provided too.

In first exam submodel we can observe these three subsections:

1. a 2 way by chance decision box that determines whether an entity requires an auxiliary person to reach the room where the examination will be made
2. the set of boxes that initialize the cycle which will be repeated for a parametric number (LoopNumber

variable) of times, each one with a duration that is assigned in the submodel according to entity tipe and colour code and to a specific gamma duration set by function:

$GAMM(betaFirstExam, RedTraumaExamDuration)$ that will be divided in duration/ LoopNumber subdurations

3. the cycle that will be repeated for LoopNumber times blocking correct resources

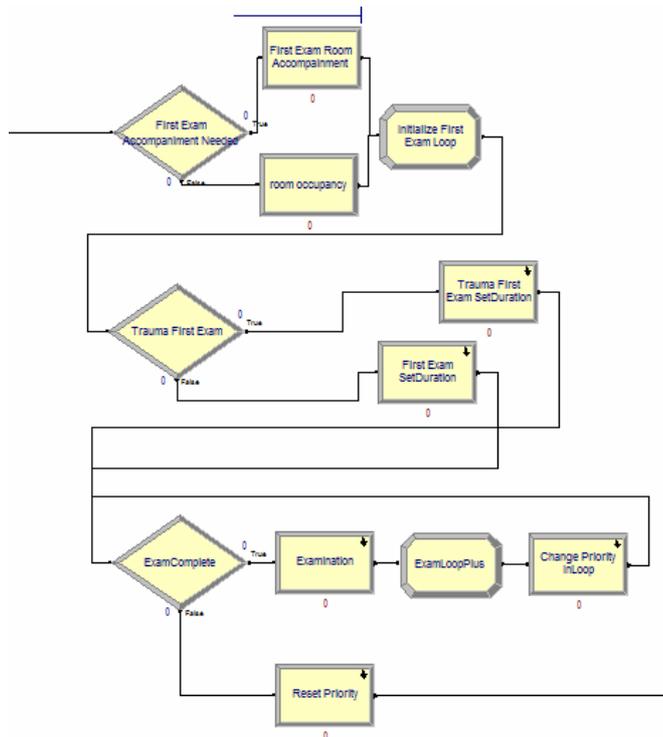


Figure 8: first examination simulation schema

The way used to manage in the right way the priority of the entities during examinations (see figure 9) is ruled by the assignment of particular values to the attribute *patientpriority* for each entity.

According to entity colour code a particular value (predefined by a variable) is set at its arrival:

RedPriority	10
YellowPriority	40
GreenPriority	50
WhitePriority	60
BluePriority	46

Remember that at lower priority value corresponds to higher priority. When entity enters in the loop simulating interruptible examination process, its patientpriority value it's lowered, so te process works trying to complete examination already started before beginning a new one.

RedPriorityLoop	5
YellowPriorityLoop	15
GreenPriorityLoop	25
WhitePriorityLoop	35
BluePriorityLoop	24

Note that only red code patients can interrupt other codes loop (other priority values are lower than the smaller priority when the entity is not looping).

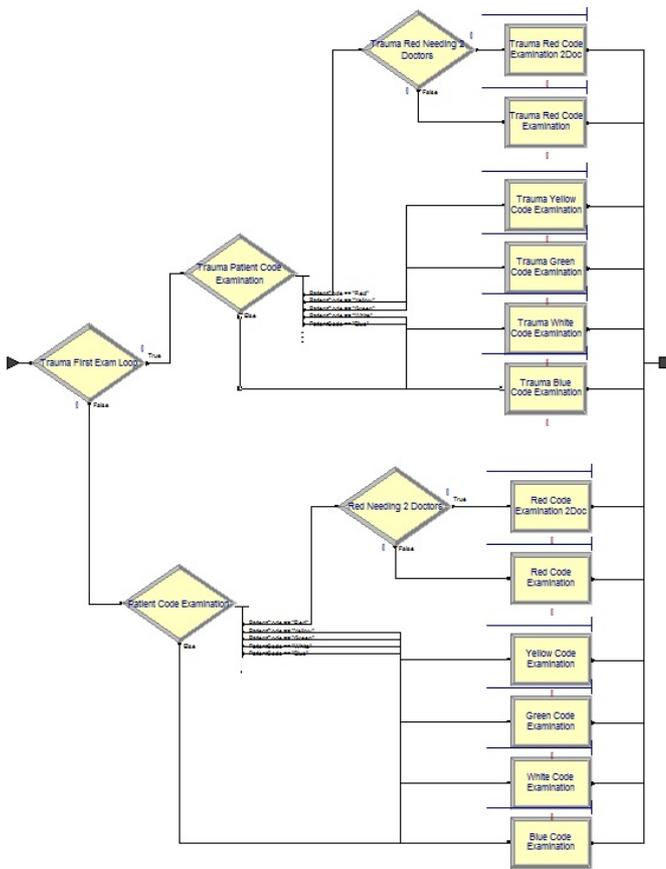


Figure 9: First Examination resources submodel

In the fourth part short term possible admission to the short term room is reproduced by means of a submodel, together with the required third examination.

The employed resources, rooms, doctors, nurses and auxiliary people are listed in separate sets, which report the amount of currently idle elements; such amounts are varied according to duties, time and assistance requests, which are satisfied according to current patient's priority or pre-emption.

Parameters reporting mean and variance of examinations duration are obtained by statistical inference from department data. Priorities are defined according to department operation rules.

4 SIMULATION RESULTS AND THEIR USE TO IMPROVE DEPARTMENT SERVICE QUALITY

4.1 Generalities

Arena model may be run to reproduce a comparatively long interval: one to three years are advised, with a warm-up period of two to four weeks.

The first qualitative result is the animation, which permits an approximate evaluation of queues and of general system behaviour.

From the quantitative point of view, the software permits to know:

- all values of patients' flow for every step;
- waiting times and queue lengths (minimum, average and maximum) both for single patient types and for single examinations;
- total patients' time in the department, separated for single patient types;
- resource utilization.

From the analysis of above results effectiveness and efficiency of the department service can be evaluated.

Only to give some examples:

- from the patients' point of view, waiting time for red code first examination shall be zero, waiting time for yellow codes first examination shall be short, total time for green and white codes shall be acceptable, etc.;
- from the operators' point of view, mean busy times and work pikes shall be acceptable, otherwise the performance decreases;
- from the administrators' point of view, the costs shall be bearable.

From the evaluation of department performance useful changes in resources at disposition, turns of duty and working rules can be suggested.

4.2 Adaptation of general model to a specific department

This is obtained without any change of model structure, but only by means of parameter setting.

More precisely we have to set:

- arrival schedulers (possibly some may be set to zero), whose parameters are obtained by past time statistics;
- examinations' duration distributions, whose parameters are obtained by past time statistics;
- resources' schedulers, whose parameters are obtained by department service rules;
- attributes ruling priorities, preemption, and resources' employment, obtained by department service rules.

4.3 Model applications and related results

The model was applied to several different emergency departments; for every application the model was validated by comparison with current true data. From the examination of actual performance, critical aspects were evidenced, and changes in the service management were tested by resetting some of the model parameters; the effects of such changes, in terms of new service performance, were reported to department managers in order to evaluate them and consider possible implementations.

The model was applied to nine emergency departments, more precisely:

- four emergency departments of city hospitals, with workloads between 90,000 and 150,000 patients per year and no ambulance management; two of them were internally divided in two areas for traumatic and medical assistance;

one of them was tested also for the case of maxi-emergencies;

- four emergency departments of small city hospitals, serving a large geographical area including many small urban nuclei, with workloads between 20,000 and 70,000 patients per year and internal ambulance management; one of them was tested also for the case of maxi-emergencies;

- one emergency department of a specialized pediatric hospital with a workload of 20,000 patients per year;

- two urgent assistance centres, able to treat only green and white codes, while red and yellow codes are transferred to a larger neighbour hospital by an ambulance under the assistance of a doctor (for red codes) or of a nurse (for yellow codes);

For all examined departments we found that the assistance to red codes was perfect, while for some of them the waiting time for lower codes was too long, and the workload for medical personnel was very high for some intervals of the day; in such cases a different setting of turns of duty with a light increase of employed personnel was suggested. In one case only an auxiliary person needed to be added to solve patients' transportation problems.

5 CONCLUSION

A generalized, adaptable model of almost all existing emergency department has been built and implemented on the computer by means of an easily readable and usable tool running on personal computers. The model can be adapted to simulate any emergency department after the only parameter setting without requiring structural modifications. Many applications proved model effectiveness.

REFERENCES

- [1] Aboueljineane L, Sahin E, Jemai Z (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*: 66 (4):734-750.
- [2] Albin SL, Wassertheil-Smoller S, Jacobson S, and Bell B (1975). Evaluation of emergency room triage performed by nurses, *Am J Public Health* 65(10):1063-1068.
- [3] Altinel K, Ulas E (1996). Simulation modeling for emergency bed requirement planning. *Annals of Operations Research* 67 (1):183-210.
- [4] Ceglowski R, Churilov L, Wasserthiel J (2005). Facilitating Decision Support in Hospital Emergency Departments: A Process Oriented Perspective. *Proceedings ECIS 2005*, Electronic support.
- [5] Ceglowski R, Churilov L, Wasserthiel J (2007). Combined Data Mining and Discrete Event Simulation for A Value Added View of A Hospital Emergency Department. *Journal of the Operational Research Society* 58(2):246-254
- [6] Clark TD, Waring CW (1987). A simulation approach to analysis of emergency services and trauma centre management. *Proceedings of the 1987 Winter Simulation Conference*, eds Thesen A, Grant H, Kelton WD: 925-931.

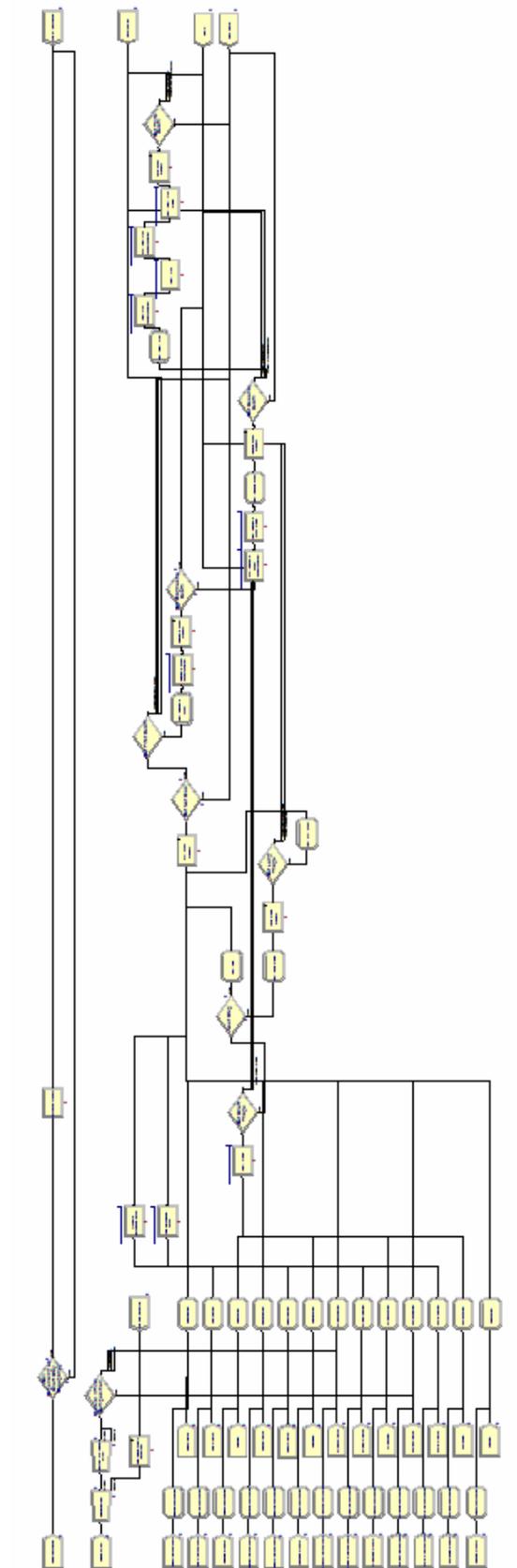


Figure 10: Arena Full simulation model

- [7] Connelly LG, Bair AE (2004). Discrete Event Simulation of Emergency Department Activity: A Platform for System-Level Operations Research. *Academy Emergency Medicine* 11: 1177-1185.
- [8] Considine J, LeVasseur SA, Charles A (2001). Consistency of Triage in Victoria's Emergency Departments: Guidelines for Triage Education and Practice. Report to the Victorian Department of Human Services. Monash Institute of Health Services Research. Australia.
- [9] Draeger MA(1992). An Emergency Department Simulation Model Used To Evaluate Alternative Nurse Staffing And Patient Population Scenarios. Proceedings of 1992 Winter Simulation Conference, eds. Swain JJ, Goldsman D, Crain RC and Wilson JR: 1057-1064.
- [10] Duguay C, Chetouane F (2007). Modelling and Improving Emergency Department Systems Using Discrete Event Simulation. *Simulation* 83 (4):311-320.
- [11] Elkum N, Fahim M, Shoukri M, Al-Madoug A (2009). Which patients wait longer to be seen and when. A waiting time study in the emergency department. *Eastern Mediterranean Health Journal* 15(2):416-424.
- [12] Evans GW, Gor TB, Unger E (1996). A Simulation Model For Evaluating Personnel Schedules In A Hospital Emergency Department. Proceedings of the 1996 Winter Simulation Conference, eds. Charnes JM, Morrice DJ, Brunner DT and Swain JJ:1205-1209
- [13] Facchin P, Rizzato E, Romanin-Jacur G (2010) Emergency department generalized flexible simulation model. IEEE Workshop on Health Care Management, Venice, Italy, Electronic support.
- [14] George S, Read S, Westlake L, Williams B, Pritty P, Fraser-Moodie A(1993). Nurse triage in theory and in practice, *Arch. Emerg. Med.* 10 (3):220–228.
- [15] Glaa B, Hammadi S, Tahon C (2006). Modeling the emergency path handling and emergency department simulation, 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan:4585-4590
- [16] Green LV, Soares J, Giglio JF, Green RA (2006). Using Queuing Theory to Increase the Effectiveness of Emergency Department Provider Staffing, *Academic Emergency Medicine*:1-8
- [17] Komashine A, Mousavi A (2005). Modeling emergency departments using discrete simulation techniques. Proceedings of the 2005 Winter Simulation Conference, eds. Kuhl ME, Syteiger NM, Armstrong FB and Joines JA: 2681-2685.
- [18] Kozan E, Diefenbach M (2008). Hospital Emergency Department Simulation for Resource Analysis. *IEMS* 7(2):133-142.
- [19] Laskowski M, Mukhi S (2009). Agent-based simulation of emergency departments with patient diversion, *Electronic Healthcare 2008 LNICST1*: 25-37.
- [20] Lopez-Valcarel BG (1994). Evolution of alternative functional designs in an emergency department by means of simulation. *Simulation* 63(1):20-28.
- [21] Mackay M., Qin S., Clissold A., Hakendorf P., Ben-Tovim D., McDonnell G. (2013) Patient flow simulation modelling – an approach conducive to multi-disciplinary collaboration towards hospital capacity management. 20th International Congress on Modelling and Simulation, Adelaide, Australia, www.mssanz.org.au/modsim2013.
- [22] McDonald L, Butterworth T, Yates DW (1995). Triage: a literature review 1985-1993. *Accident and Emergency Nursing* 3 (4): 201-207.
- [23] Pezij JW (2007). Testing scenarios in a Simulation Model of the Emergency Department, University of Twente Student Theses, Twente, The Netherlands.
- [24] Rowe JA (1992). Triage assessment tool. *Journal of emergency nursing* 18(6):540-544.
- [25] Romanin-Jacur G, Hospital Maxi-Emergency Protocol Testing by A Double Dynamics Simulation Model (2005). *Modelling and Simulation 2005*, eds. Feliz Teixeira JM and Carvalho Brito AE, EUROSIS-ETI, Ghent, Belgium:153-156.
- [26] Ruohonen T, Neittaanmaki P, Teittinem J (2006). Simulation Model for Improving the Operation of the Emergency Department of Special Health Care. Proceedings of 2006 Winter Simulation Conference, eds. Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM and Fujimoto RM:453-458.
- [27] Saunders CE, Makens PK, Leblanc LJ (1989). Modelling emergency department operations using advanced computer simulation systems. *Ann. Emergency Medicine* 18(2):134-40.
- [28] Solomon M, Jacobson J, Grigsby E, Pennbridge J, Le Q, Singleton O (2003). A Discrete-Event Simulation Model of Inpatient and Emergency Services in Los Angeles County. *Abstracts of Academy Health Meetings* 20:670.
- [29] Su S, Shih CL. (2002) Resource reallocation in an emergency medical service system using computer simulation. *Am J Emerg Med.* 20(7):627-34.
- [30] Su S, Shih CL. (2003) Modeling an emergency medical services system using computer simulation. *Int J Med Inform.* 72(1-3):57-72. 23 casi in Cina con suggerimenti, studio dei servizi
- [31] Wang T, Guinet A, Besombes B (2008). Simulation modeling of emergency service with the impact of inpatient bed resource. *International Conference on Information Systems, Logistics and Supply Chain, Madison WI, U.S.A.*, Electronic support.
- [32] Yeh JY, Lin WS (2007). Using Simulation Technique and Genetic Algorithm to Improve the Quality Care of Hospital Emergency Department. *Expert Systems with Applications* 32:1073-1083.
- [33] Zilm F (2004). Estimating Emergency Service Treatment Bed Needs. *Ambulatory Care Management* 27(3):215-223

Modelling, Simulation and Control of Technological Processes

NONLINEAR CONTROL OF A SHELL AND TUBE HEAT EXCHANGER

Petr Dostál^{1,2}, Vladimír Bobál^{1,2}, and Jiří Vojtěšek²

¹Centre of Polymer Systems, University Institute, Tomas Bata University in Zlin,
Nad Ovcirnou 3685, 760 01 Zlin, Czech Republic.

²Department of Process Control, Faculty of Applied Informatics, Tomas Bata University in Zlin,
Nad Stranemi 4511, 760 05 Zlin, Czech Republic
{dostalp;bobal;vojtesek}@fai.utb.cz

KEYWORDS

Tubular heat exchanger, nonlinear model, steady-state characteristics, external linear model, parameter estimation, polynomial approach, control simulation.

ABSTRACT

The paper deals with design and simulation of nonlinear adaptive control of a shell and tube heat exchanger. The method is based on factorization of the controller on a nonlinear static part and an adaptive linear dynamic part. The nonlinear static part is derived using inversion and subsequent exponential approximation of simulated or measured steady-state characteristics of the exchanger. The linear dynamic part is then obtained from an external linear model of nonlinear elements of the closed-loop. The parameters of the external linear model are recursively estimated via a corresponding delta model. The control law in the 1DOF and 2DOF control system structures is derived using the polynomial approach.

INTRODUCTION

Heat exchangers are an essential part of many technologies in energy and chemical industry, polymer manufacturing, petroleum refineries, and many others. By construction, heat exchangers can be classified into exchangers with direct contact, various types of plate exchangers, and, shell and tube heat exchangers (STHEs), see, e.g. (Smith 2005; Hewitt et al. 1994; Incropera et al. 2011).

As known, STHEs are most common types of heat exchangers. From the system theory, they belong to a class of nonlinear distributed parameter systems with mathematical models in the form of nonlinear partial differential equations. Modelling and simulation of such processes are described in many publications, e.g. in (Luyben 1989; Corriou 2004; Babu 2004; Ogunnaike and Rao 1994). As known, these processes can be hardly controllable by conventional methods that can lead to control of a poor quality. In this case, some advanced control methods should be used such as adaptive, predictive, optimal or nonlinear control, and some others. Obviously, control design always requires a preliminary steady-state and dynamic analysis of the process by simulation tools. Some methods of numerical mathematics used to build simulation models can be

found e.g. in (Nevriva et al. 2009; Cook 2002).

The aim of the paper is an application of nonlinear control and subsequent control simulation of a simple type of the shell and tube heat exchanger. The control strategy is based on the idea of factorization of the controller on a nonlinear static part (NSP) and an adaptive linear dynamic part (LDP). Similar approaches can be found e.g. in (Chen et al. 2006; Dostál et al. 2011b). The nonlinear static part is obtained from simulated or measured steady-state characteristic of the STHE, its inversion, exponential approximation, and, subsequently, its differentiation. On behalf of development of the linear part, the NSP including the nonlinear model of the STHE is approximated by a continuous-time external linear model (CT ELM). For the CT ELM parameter estimation, an external delta model with the same structure as the CT model is used. The basics of delta models have been described e.g. in (Mukhopadhyay et al. 1992; Garnier and Wang 2008). Although delta models belong into discrete models, they do not have such disadvantageous properties connected with shortening of a sampling period as discrete z -models. In addition, parameters of delta models can directly be estimated from sampled signals. Moreover, it can be easily proved that these parameters converge to parameters of CT models for a sufficiently small sampling period (compared to the dynamics of the controlled process), as shown in (Stericker and Sinha 1993; Dostál et al. 2004).

The 1DOF and the 2DOF control system structures are considered. In the first case, the control system includes only a feedback controller, in the second case, the controller consist of a feedback and a feedforward part. Such structures were described and applied e.g. in (Dostál et al. 2011a; Grimble 1993). Then, resulting CT controllers are derived using the polynomial approach and the pole placement method (Kučera 1993; Mikleš and Fikar 2004).

The simulations are performed on a nonlinear model of the STHE.

MODEL OF THE STHE

Consider an ideal plug-flow shell and tube heat exchanger in the fluid phase and with the counterflow cooling. The fluid flowing in tubes is cooled by a fluid flowing in the shell as shown in Fig. 1. Heat losses and heat conduction along the metal walls of tubes are

assumed to be negligible, but dynamics of the metal walls of tubes are significant. All densities, heat capacities, and heat transfer coefficients are assumed to be constant. Under above assumptions, the STHE model can be described by three partial differential equations (PDEs) in the form

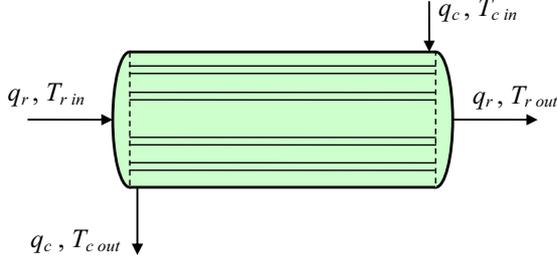


Figure 1: Shell and Tube Heat Exchanger.

$$\frac{\partial T_r}{\partial t} + v_r \frac{\partial T_r}{\partial z} = c_1 (T_w - T_r) \quad (1)$$

$$\frac{\partial T_w}{\partial t} = c_2 (T_r - T_w) + c_3 (T_c - T_w) \quad (2)$$

$$\frac{\partial T_c}{\partial t} - v_c \frac{\partial T_c}{\partial z} = c_4 (T_w - T_c) \quad (3)$$

with initial conditions

$$T_r(z, 0) = T_r^s(z), \quad T_w(z, 0) = T_w^s(z), \quad T_c(z, 0) = T_c^s(z)$$

and boundary conditions

$$T_r(0, t) = T_{r0}(t), \quad T_c(L, t) = T_{cL}(t).$$

The parameters in (1) – (3) are

$$\begin{aligned} v_r &= \frac{4q_r}{n_1 \pi d_1^2}, \quad v_c = \frac{4q_c}{\pi (d_3^2 - n_1 d_1^2)} \\ c_1 &= \frac{4\alpha_1}{d_1 \rho_r c_{pr}}, \quad c_2 = \frac{4d_1 \alpha_1}{(d_2^2 - d_1^2) \rho_w c_{pw}} \\ c_3 &= \frac{4d_2 \alpha_2}{(d_2^2 - d_1^2) \rho_w c_{pw}}, \quad c_4 = \frac{4n_1 d_2 \alpha_2}{(d_3^2 - n_1 d_2^2) \rho_c c_{pc}} \end{aligned} \quad (4)$$

where t stands for the time, z for the axial space variable, T for temperatures, q for flow of fluids, v for fluid flow velocities, d_1 for inner diameter of the tube, d_2 for outer diameter of the tube, d_3 for diameter of the shell, ρ for densities, c_p for specific heat capacities, α for heat transfer coefficients, n_1 is the number of tubes and L is the length of tubes. Subscripts denoted r describe the refrigerated fluid (RF), w the metal walls of tubes, c the cooling fluid (CF), and the superscript s steady-state values.

From the system engineering point of view, $T_r(L, t) = T_{rout}$ and $T_c(0, t) = T_{cout}$ are the output variables, and, $q_r(t)$, $q_c(t)$, $T_{r0}(t)$ and $T_{cL}(t)$ are the input variables. For the control purposes, the output temperature of the refrigerated fluid $T_r(L, t) = T_{rout}(t)$ is considered as the controlled output, and, the coolant

flow $q_c(t)$ as the control input, while other inputs can enter into the process as disturbances. The parameter and steady-state input values with their correspondent units are given in Table 1.

Table 1: Parameters and Steady-State Inputs

$L = 8$ m	$n_1 = 1100$
$d_1 = 0.022$ m	$d_2 = 0.024$ m
$d_3 = 1$ m	
$\rho_r = 985$ kg/m ³	$c_{pr} = 4.05$ kJ/kg K
$\rho_w = 7800$ kg/m ³	$c_{pw} = 0.71$ kJ/kg K
$\rho_c = 998$ kg/m ³	$c_{pc} = 4.18$ kJ/kg K
$\alpha_1 = 5.8$ kJ/m ² s K	$\alpha_2 = 3.6$ kJ/m ² s K
$T_{r0}^s = 373$ K	$T_{cL}^s = 293$ K
$q_r^s = 0.1$ m ³ /s	$q_c^s = 0.09$ m ³ /s

COMPUTATION MODELS

For computation of both steady-state and dynamic characteristics, the finite differences method is employed. The procedure is based on substitution of the space interval $z \in \langle 0, L \rangle$ by a set of discrete node points $\{z_i\}$ for $i = 1, \dots, n$, and, subsequently, by approximation of derivatives with respect to the space variable in each node point by finite differences. Two types of finite differences are applied, either the backward or the forward finite difference.

Dynamic Model

Applying the finite differences method, PDEs (1) – (3) are approximated by a set of ODEs in the form

$$\frac{dT_r(i)}{dt} = -\left(\frac{v_r}{h} + c_1\right)T_r(i) + \frac{v_r}{h}T_r(i-1) + c_1T_w(i) \quad (5)$$

$$\frac{dT_w(i)}{dt} = c_2[T_r(i) - T_w(i)] + c_3[T_c(i) - T_w(i)] \quad (6)$$

$$\frac{dT_c(j)}{dt} = -\left(\frac{v_c}{h} + c_4\right)T_c(j) + \frac{v_c}{h}T_c(j+1) + c_4T_w(j) \quad (7)$$

for $i = 1, \dots, n$, $j = n - i + 1$, and, with initial conditions $T_r(i, 0) = T_r^s(i)$, $T_w(i, 0) = T_w^s(i)$ and $T_c(i, 0) = T_c^s(i)$ for $i = 1, \dots, n$. In (5) – (7), h is the diskretization step.

The boundary conditions enter into Eqs. (5) – (7) for $i = 1$.

Here, the controlled output is computed as

$$T_{rout}(t) = T_r(n, t) \quad (8)$$

Steady-State Model

Computation of steady-state characteristics is necessary not only for a steady-state analysis but the steady state values also constitute initial conditions in ODEs (5) – (7). The steady-state model can simply be derived

equating the time derivatives in (5) – (7) to zero. Then, the steady-state characteristics can be computed by an iterative method.

Steady-State Characteristics

The dependence of the RF output temperature on the coolant flow in the steady-state is in Fig. 2. In subsequent control simulations, the operating interval for q_c has been determined as

$$q_c^{min} \leq q_c(t) \leq q_c^{max}. \quad (9)$$

With regard to the purposes of a latter approximation of the steady-state characteristics, the values $q_c^L < q_c^{min}$ and $q_c^U > q_c^{max}$ are established that denote the lower and upper bound of q_c^s used for the approximation. Their values together with values in (9), and, to them corresponding temperatures are in Table 2.

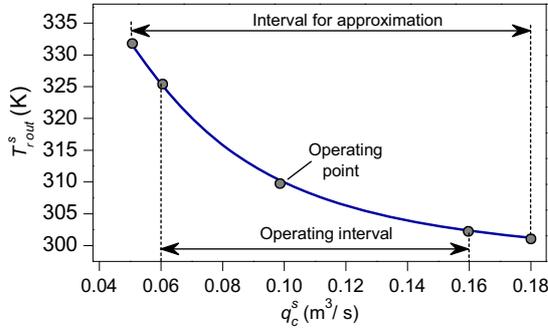


Figure 2: Dependence of the RF Output Temperature on the Coolant Flow Rate.

Table 2: Input and Output Values Used in Approximation.

$q_c^L = 0.05 \text{ m}^3/\text{s}$	$q_c^U = 0.18 \text{ m}^3/\text{s}$
$T_{rout}^U = 332.09 \text{ K}$	$T_{rout}^L = 301.21 \text{ K}$
$q_c^{min} = 0.06 \text{ m}^3/\text{s}$	$q_c^{max} = 0.16 \text{ m}^3/\text{s}$
$T_r^{max} = 325.49$	$T_r^{min} = 302.35$

CONTROLLER DESIGN

As previously introduced, the controller consist of a nonlinear static part and a linear dynamic part as shown in Fig. 3.

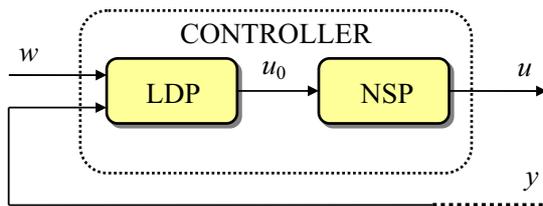


Figure 3: Controller Scheme.

Here, the control input and the controlled output variables are considered in the form

$$u(t) = \Delta q_c(t) = q_c(t) - q_c^s, \quad y(t) = T_{rout}(t) - T_{rout}^s \quad (10)$$

where ($q_c^s = 0.9, T_{rout}^s = 312.53$) represents an operating point around which the changes take place during the control.

The LDP creates a linear dynamic relation which represents a difference of the RF output temperature adequate to its desired value. Then, the NSP generates a static nonlinear relation between u_0 and a corresponding increment (decrement) of the coolant flow rate.

Nonlinear Static Part of the Controller

The NSP derivation appears from a simulated or measured steady-state characteristics. The coordinates on the graph axis are defined as

$$\xi = \frac{q_c^s - q_c^L}{q_c^U - q_c^L}, \quad \psi = \frac{T_{rout}^s - T_{rout}^L}{T_{rout}^U - T_{rout}^L} \quad (11)$$

where

$$q_c^L \leq q_c^s \leq q_c^U. \quad (12)$$

In term of the practice, it can be supposed that the measured data will be affected by measurement errors. The simulated steady-state characteristics that corresponds to reality is shown in Fig. 4.

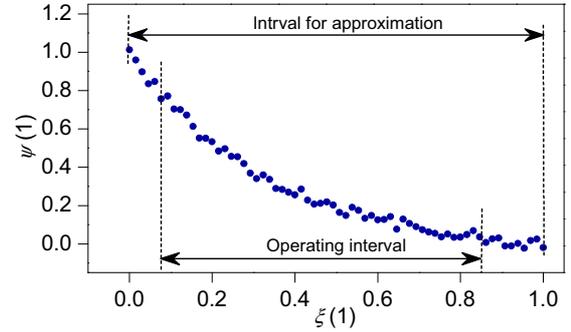


Figure 4: Steady-State Characteristics in Presence of Disturbance.

Changing the axis, the inverse of this characteristic can be approximated by a function from the ring of polynomial, exponential, rational, eventually, by other type functions. Here, the second order exponential approximate function has been found in the form

$$\xi = -1.3 + 0.619 e^{-4.505\psi} + 1.606 e^{-0.213\psi}. \quad (13)$$

The inverse characteristic together with its approximation is in Fig. 5. Now, a difference of the coolant flow rate $u(t) = \Delta q_c(t)$ in the output of the NSP can be computed for each T_{rout} as

$$u(t) = \Delta q_c(t) = \frac{q_c^U - q_c^L}{T_{rout}^U - T_{rout}^L} \left(\frac{d\xi}{d\psi} \right)_{\psi(T_{rout})} u_0(t) \quad (14)$$

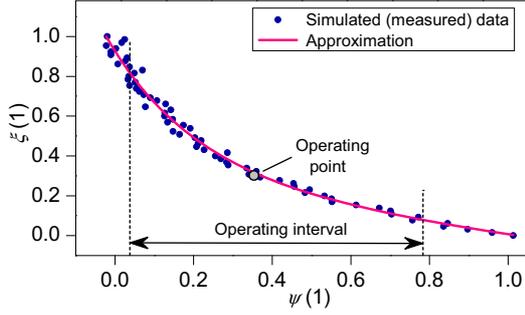


Figure 5: Approximation of the Inverse Characteristics.

where the derivative of ξ with respect to ψ takes the form

$$\frac{d\xi}{d\psi} = -2.789e^{-4.505\psi} - 0.342e^{-0.213\psi}. \quad (15)$$

Its plot is shown in Fig. 6.

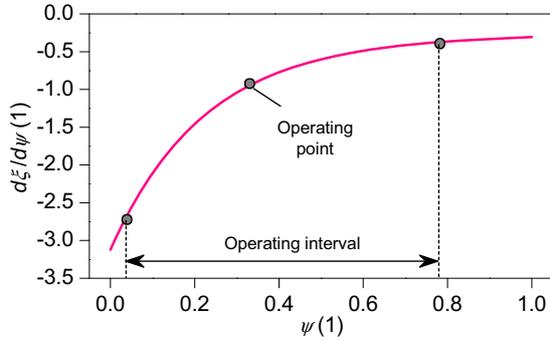


Figure 6: Derivative of Inverse Characteristics.

CT and Delta External Linear Model

The nonlinear component (NC) of the closed-loop consisting of the NSP of the controller and the STHE nonlinear model is shown in Fig. 7.

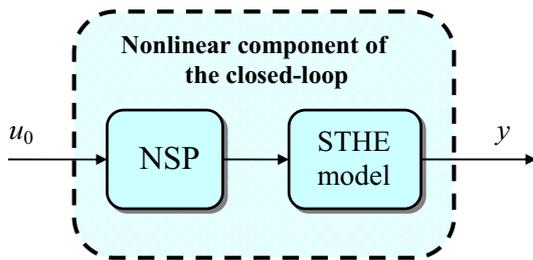


Figure 7: Nonlinear Component of the Closed-Loop.

The CT external linear model of this nonlinear component is chosen on the basis of step responses simulated around the above defined operating point. Step responses are shown in Fig. 8.

Taking into account profiles of curves in Fig. 8 with zero derivatives in $t = 0$, the second order CT ELM has been chosen in the form of the second order linear differential equation

$$\ddot{y}(t) + a_1 \dot{y}(t) + a_0 y(t) = b_0 u(t) \quad (16)$$

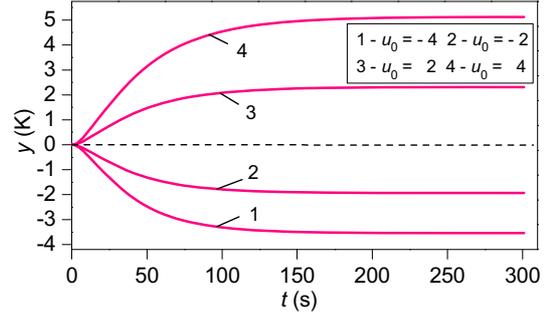


Figure 8: NC Output Step Responses.

and, in the complex domain, as the transfer function

$$G(s) = \frac{b_0}{s^2 + a_1 s + a_0}. \quad (17)$$

Establishing the δ -operator

$$\delta = \frac{q-1}{T_0} \quad (18)$$

where q is the forward shift operator and T_0 is the sampling period, the delta ELM corresponding to (17) takes the form

$$\delta^2 y(t') + a'_1 \delta y(t') + a'_0 y(t') = b'_0 u(t') \quad (19)$$

where t' is the discrete time. When the sampling period is shortened, the delta operator approaches the derivative operator, and, the estimated parameters a', b' reach the parameters a, b of the CT model.

Delta Model Parameter Estimation

Substituting $t' = k-2$, equation (19) can be rewritten to the form

$$\delta^2 y(k-2) + a'_1 \delta y(k-2) + a'_0 y(k-2) = b'_0 u(k-2). \quad (20)$$

Establishing the regression vector

$$\Theta_\delta^T(k-1) = (-\delta y(k-2) \quad -y(k-2) \quad u(k-2)) \quad (21)$$

where

$$\delta y(k-2) = \frac{y(k-1) - y(k-2)}{T_0} \quad (22)$$

the vector of delta model parameters

$$\Theta_\delta^T(k) = (a'_1 \quad a'_0 \quad b'_0) \quad (23)$$

is recursively estimated by the least squares method with exponential and directional forgetting from the ARX model, e.g. (Bobál et al. 2005).

$$\delta^2 y(k-2) = \Theta_\delta^T(k) \Phi_\delta(k-1) + \varepsilon(k) \quad (24)$$

where

$$\delta^2 y(k-2) = \frac{y(k) - 2y(k-1) + y(k-2)}{T_0^2}. \quad (25)$$

Linear Dynamic Part of the Controller

In the control simulations, the 1DOF and 2DOF control system structures are considered. While the 1DOF structure includes only the feedback controller Q , a controller in the 2DOF structure consist of the feedback part Q and the feedforward part R as shown in Figs. 9 and 10. In both figures, w denotes the reference signal, y the controlled output and u_0 the input to the ELM. The reference w and the disturbance v are considered as step

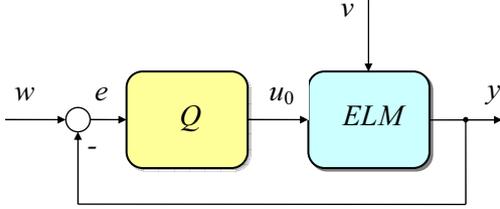


Figure 9: 1DOF Control System Structure.

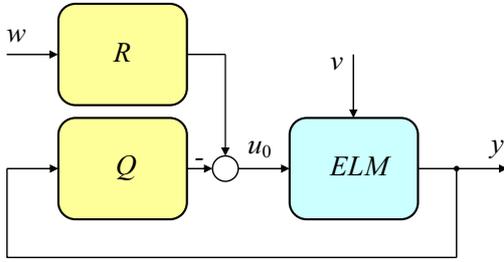


Figure 10: 2DOF Control System Structure.

functions with transforms

$$W(s) = \frac{w_0}{s}, \quad V(s) = \frac{v_0}{s}. \quad (26)$$

The controller transfer functions in both structures are considered as

$$Q(s) = \frac{q(s)}{p(s)}, \quad R(s) = \frac{r(s)}{p(s)} \quad (27)$$

where q , r and p are coprime polynomials in s fulfilling conditions of properness $\deg r \leq \deg p$ and $\deg q \leq \deg p$. For both step input signals, the polynomial p takes the form $p(s) = s\tilde{p}(s)$.

It is well known from the algebraic control theory that controllers ensuring stability of the control system result in the polynomial ring from a solution of the polynomial equation

$$a(s)p(s) + b(s)q(s) = d(s) \quad (28)$$

in the 1DOF structure, and, from a couple of polynomial equations

$$a(s)p(s) + b(s)q(s) = d(s) \quad (29)$$

$$st(s) + b(s)r(s) = d(s) \quad (30)$$

in the 2DOF structure. In both cases, the polynomial d on their right sides is a stable polynomial.

Then, the controller's transfer functions take forms

$$Q(s) = \frac{q(s)}{s\tilde{p}(s)} = \frac{q_2s^2 + q_1s + q_0}{s(s+p_0)} \quad (31)$$

$$R(s) = \frac{r(s)}{s\tilde{p}(s)} = \frac{r_0}{s(s+p_0)} \quad (32)$$

where $Q(s)$ is the same for both structures. Moreover, the equality $r_0 = q_0$ can easily be obtained.

The controller parameters then follow from solutions of polynomial equations (29) and (30) and depend upon coefficients of the polynomial d . In this paper, the polynomial d with roots determining the closed-loop poles is chosen as

$$d(s) = n(s)(s + \alpha)^2 \quad (33)$$

where n is a stable polynomial obtained by spectral factorization

$$a^*(s)a(s) = n^*(s)n(s) \quad (34)$$

and α is a selectable parameter that can usually be chosen by way of simulation experiments. Note that a choice of d in the form (33) provides the control of a good quality for aperiodic controlled processes. The polynomial n has the form

$$n(s) = s^2 + n_1s + n_0 \quad (35)$$

with coefficients

$$n_0 = \sqrt{a_0^2}, \quad n_1 = \sqrt{a_1^2 + 2n_0 - 2a_0}. \quad (36)$$

Then, the controller parameters can be obtained from solution of the matrix equation

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ a_1 & b_0 & 0 & 0 \\ a_0 & 0 & b_0 & 0 \\ 0 & 0 & 0 & b_0 \end{pmatrix} \cdot \begin{pmatrix} p_0 \\ q_2 \\ q_1 \\ q_0 \end{pmatrix} = \begin{pmatrix} d_3 - a_1 \\ d_2 - a_0 \\ d_1 \\ d_0 \end{pmatrix} \quad (37)$$

where

$$\begin{pmatrix} d_3 = n_1 + 2\alpha, & d_2 = 2\alpha n_1 + n_0 + \alpha^2 \\ d_1 = 2\alpha n_0 + \alpha^2 n_1, & d_0 = \alpha^2 n_0 \end{pmatrix}. \quad (38)$$

Evidently, the controller parameters can be adjusted by the selectable parameter α .

SIMULATION RESULTS

All control simulations were performed around the above defined operating point.

For the start (adaptation phase 15 min), a P controller with experimentally tuned small gain was used in all simulations.

An effect of the parameter α on the controlled output and the control input responses in the 1DOF structure is presented in Figs. 11 and 12. While a difference between controlled outputs for selected values α is not very significant, an increasing α results in higher changes of the control input. Both signals significantly respond to further increase of the value α as it can be seen in Figs. 13 and 14 for the 1DOF structure. There, the controlled output exhibits higher overshoots, and,

the control input oscillates between constraints. Therefore, the selection of an appropriate value α is very important especially in control of a real process. The controlled output and the control input in both control system structures for a selected value α are compared in Figs. 15 and 16. There, a difference in controlled outputs is minimal, but the control input in the 2DOF structure shows smaller changes. The big difference in the use of both structures for the higher α is but to see in Figs. 13 and 14 where the results from the 1DOF structure are unacceptable while the 2DOF structure provides good control responses.

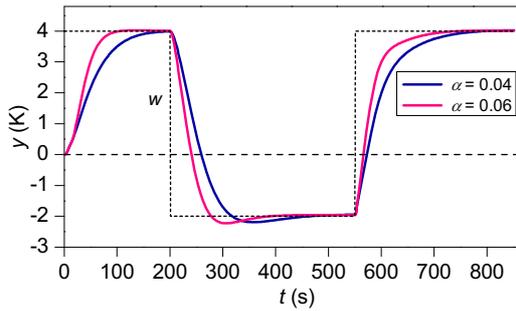


Figure 11: Controlled Output Responses for Various α in the 1DOF Control Structure.

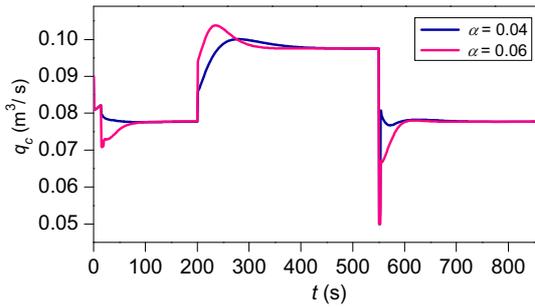


Figure 12: Coolant Flow Responses for Various α in the 1DOF Control Structure.

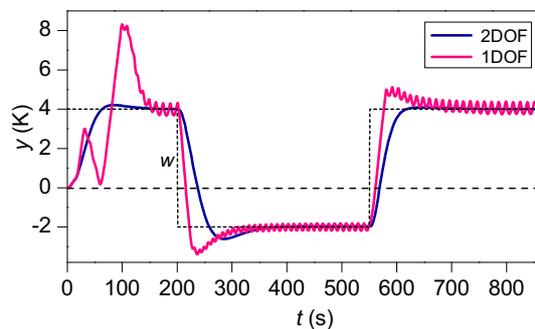


Figure 13: Controlled Output Responses in the 1DOF and 2DOF Control Structures ($\alpha = 0.08$).

CONCLUSIONS

The paper dealt with design and simulation of nonlinear adaptive control of a shell and tube heat exchanger. The control strategy is based on factorization of a

controller into the linear and the nonlinear part. The design of the controller nonlinear part employs steady-state characteristics of the process and their additional

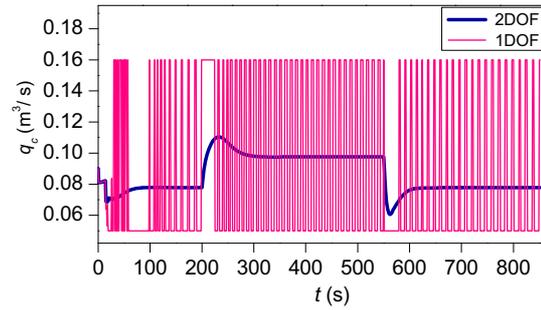


Figure 14: Coolant Flow Responses in the 1DOF and 2DOF Control Structures ($\alpha = 0.08$).

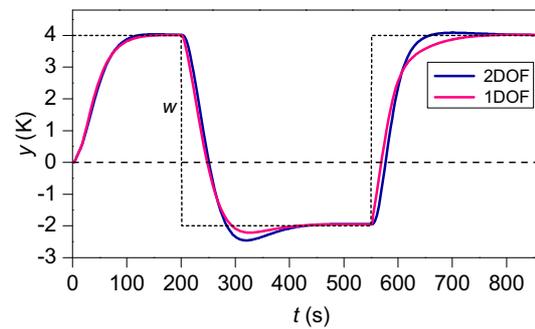


Figure 15: Controlled Output Responses in the 1DOF and 2DOF Control Structures ($\alpha = 0.05$).

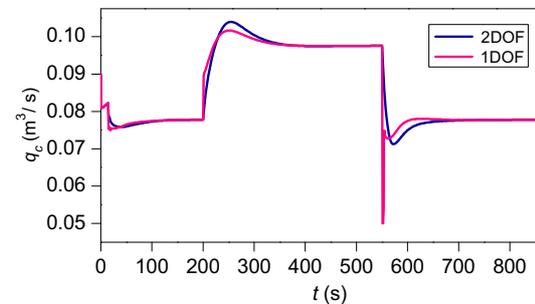


Figure 16: Coolant Flow Responses in the 1DOF and 2DOF Control Structures ($\alpha = 0.05$).

modifications. Then, the nonlinear part of the closed-loop is approximated by a continuous time external linear model with parameters obtained via recursive parameter estimation of a corresponding delta model. The resulting linear part of the controller is considered for both 1DOF and 2DOF control system structures and derived using the polynomial approach. The simulation results demonstrate improved usability of the 2DOF structure especially in terms of the control input characteristics.

ACKNOWLEDGEMENT

This article was written with support of Operational Program Research and Development for Innovations co-funded by the European Regional Development Fund

(ERDF) and national budget of Czech Republic, within the framework of project Centre of Polymer Systems (reg. number: CZ.1.05/2.1.00/03.0111).

REFERENCES

- Babu, B.V. 2004. *Process plant simulation*. Oxford University Press, New York.
- Bobál, V., J. Böhm, J. Fessl, and J. Macháček. 2005. *Digital self-tuning controllers*, Springer Verlag, Berlin, 2005.
- Chen, C.-T., Y.-C. Chuang, and C. Hwang. 2006. "A simple nonlinear control strategy for chemical processes". In: *Proceedings of 6th Asian Control Conference*, Bali, Indonesia, 64–70.
- Cook, R.D. 2002. *Applications of finite element analysis*. John Wiley and Sons, Chichester 2002.
- Corriou, J.-P. 2004. *Process control. Theory and applications*. Springer – Verlag, London.
- Dostál, P., V. Bobál, and F. Gazdoš. 2004. "Adaptive control of nonlinear processes: Continuous-time versus delta model parameter estimation". In: *Proceedings of 8th IFAC Workshop on Adaptation and Learning in Control and Signal Processing ALCOSP 04*, Yokohama, Japan, 273-278.
- Dostál, P. J. Vojtěšek, and V. Bobál. 2011a. "Simulation of the 2DOF nonlinear adaptive control of a chemical reactor". In: *Proceedings of 25th European Conference on Modelling and Simulation*, Krakow, Poland, 494-499.
- Dostál, P., V. Bobál, and F. Gazdoš. 2011b. "Simulation of nonlinear adaptive control of a continuous stirred tank reactor". *International Journal of Mathematics and Computers in Simulation*, 5, 370-377.
- Garnier, H. and L. Wang (eds.). 2008. *Identification of continuous-time models from sampled data*. Springer-Verlag, London, 2008.
- Grimble, M.J. 1993. *Robust industrial control. Optimal design approach for polynomial systems*. Prentice Hall, Englewood Cliffs.
- Hewitt, G.F., G.L. Shires, and T.R. Bott. 1994. *Process Heat Transfer*. CRC Press, Inc.
- Incropera, F.P., A. S. Lavine, and P. DeWitt. 2011. *Fundamentals of heat and mass transfer*. John Wiley, NJ.
- Kučera, V. 1993. "Diophantine equations in control – A survey". *Automatica*, 29, 1361-1375.
- Luyben, W. 1989. *Process modelling, simulation and control for chemical engineers*. McGraw-Hill, New York.
- Mikleš, J. and M. Fikar. 2004. *Process modelling, identification and control 2*, STU Press, Bratislava, Slovakia, 2004.
- Mukhopadhyay, S., A.G. Patra and G.P. Rao. 1992. "New class of discrete-time models for continuous-time systems". *International Journal of Control*, 55, 1161-1187.
- Nevriva, P., S. Oyana, and L. Vilimec. 2009. "Simulation of the heat exchangers dynamics in MATLAB & Simulink". *WSEAS Transactions on Systems and Control*, 4, 519-530.
- Ogunnaike, B.A. and W.H. Rao. 1994. *Process dynamics, modeling, and control*, Oxford University Press, New York.
- Smith, R. 2005. *Chemical process design and integration*. John Wiley and Sons, Chichester.
- Stericker, D.L. and N.K. Sinha. 1993. "Identification of continuous-time systems from samples of input-output data using the δ -operator". *Control-Theory and Advanced Technology*, 9, 113-125.

AUTHOR BIOGRAPHIES



PETR DOSTÁL studied at the Technical University of Pardubice, where he obtained his master degree in 1968 and Ph.D. degree in Technical Cybernetics in 1979. In the year 2000 he became professor in Process Control. He is now Professor in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interest are modeling and simulation of continuous-time chemical processes, polynomial methods, optimal and adaptive control.



VLADIMÍR BOBÁL was born in Slavičín, Czech Republic. He graduated in 1966 from the Brno University of Technology. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interests are adaptive control systems, system identification and CAD for self-tuning controllers.



JIRÍ VOJTĚŠEK was born in Zlín, Czech Republic in 1979. He studied at Tomas Bata University in Zlín, Czech Republic, where he received his M.Sc. degree in Automation and control in 2002. In 2007 he obtained Ph.D. degree in Technical cybernetics at Tomas Bata University in Zlín. He now works as an assistant professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are modeling and simulation of continuous-time chemical processes, polynomial methods, optimal, adaptive and nonlinear control. Dr. Vojtesek is the Chairman of the European Conference of the Modelling and Simulation (ECMS) and IPC member of the IASTED.

DIGITAL LINEAR QUADRATIC SMITH PREDICTOR

Vladimír Bobál^{1,2}, Marek Kubalčík², Petr Dostál^{1,2} and Stanislav Talaš²
Tomas Bata University in Zlín

¹Centre of Polymer Systems, University Institute

²Department of Process Control, Faculty of Applied Informatics

T. G. Masaryka 5555

760 01 Zlín

Czech Republic

E-mail: bobal@fai.utb.cz

KEYWORDS

Time-delay Systems, Smith Predictor, LQ Control, Spectral Factorization, Simulation

ABSTRACT

Time-delays (dead times) are found in many processes in industry. Time-delays are mainly caused by the time required to transport mass, energy or information, but they can also be caused by processing time or accumulation. The contribution is focused on a design of universal digital algorithm for control of great deal of processes with time-delay. This requirement is successfully satisfied with digital Smith Predictor based on Linear Quadratic (LQ) method. A minimization of the quadratic criterion is realized using spectral factorization. The designed algorithm is suitable for control of stable, unstable and non-minimum phase processes. The algorithms for control of individual processes influenced by external disturbance were verified. The program system MATLAB/SIMULINK was used for simulation of designed algorithms.

INTRODUCTION

Time-delays appear not only in industrial processes, such as thermal, chemical, metallurgical or processes of plastic and rubber materials etc., but also in other fields, including economical and biological systems. They are caused by some of the following phenomena (Normey-Rico and Camacho 2007):

- the time needed to transport mass, energy or information,
- the accumulation of time lags in a great numbers of low order systems connected in series,
- the required processing time for sensors, such as analyzers; controllers that need some time to implement a complicated control algorithms or process.

The problem of controlling time-delay processes can be solved by several control methods (e. g. using PID controllers, time-delay compensators, model predictive control techniques).

Time-delay in a process increases the difficulty of controlling it. However, the approximation of higher-order process by lower-order model with time-delay

provides simplification of the control algorithms. When high performance of the control process is desired or the relative time-delay is very large, the predictive control strategy can be successfully applied. The predictive control method includes a model of the process in the structure of the controller. The first time-delay compensation algorithm was proposed by (Smith 1957). This control algorithm known as the Smith Predictor (SP) contained a dynamic model of the time-delay process and it can be considered as the first model predictive algorithm. First versions of Smith Predictors were designed in the continuous-time modifications, see e.g. (Normey-Rico and Camacho 2007).

Although time-delay compensators appeared in the mid 1950s, their implementation with analog technique was very difficult and these were not used in industry. Because most of modern controllers are implemented on digital platforms, the discrete versions of the time-delay controllers are more suitable for time-delay compensation in industrial practice

Since 1980s digital time-delay compensators can be implemented. The digital time-delay compensators are presented e.g. in (Vogel and Edgar 1980, Palmor and Halevi 1990, Normey-Rico and Camacho 1998). Some Self-tuning Controller (STC) modifications of the digital Smith Predictors (STCSP) are designed in (Hang et al. 1989; Hang et al. 1993; Bobál et al. 2011). Two versions of the STCSP were implemented into MATLAB/SIMULINK Toolbox (Bobál et al. 2012a; Bobál et al. 2012b).

It is well known that classical analog Smith Predictor is not suitable for control of unstable processes. The designed digital LQ Smith Predictor eliminates this drawback.

The paper is organized in the following way. The problem of a control of the time-delay systems is described in Section 1. The general principle of the Smith Predictor is described in Section 2. The discretization of analogue version, principle of digital Smith Predictor and polynomial two degrees of freedom (2DOF) controller is introduced in Section 3. Primary Linear Quadratic controller of the digital Smith Predictor is proposed in Section 4. Results of simulation experiments are summed in Section 5. Section 6 concludes the paper.

DIGITAL SMITH PREDICTOR

The discrete versions of the SP and their modifications are suitable for time-delay compensation in industrial practice.

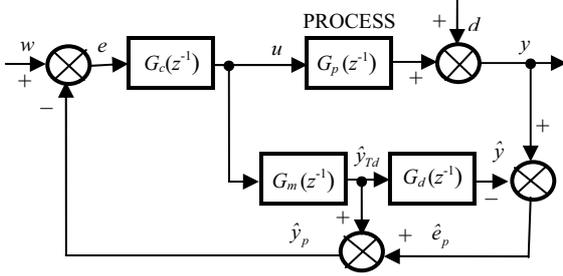


Figure 1: Block diagram of a digital Smith Predictor

The block diagram of a digital SP (see Hang, Lim, and Chong 1989; Hang, Tong, and Weng 1993) is shown in Fig. 1. The function of the digital version is similar to the classical analog version. The block $G_m(z^{-1})$ represents process dynamics without the time-delay and is used to compute an open-loop prediction. The difference between the output of the process y and the model including time delay \hat{y} is the predicted error \hat{e}_p as shown in Fig. 1, whereas e and v are the error and the measured disturbance, w is the reference signal. If there are no modelling errors or disturbances, the error between the current process output y and the model output \hat{y} will be null. Then the predictor output signal \hat{y}_p will be the time-delay-free output of the process. Under these conditions, the controller $G_c(z^{-1})$ can be tuned, at least in the nominal case, as if the process had no time-delay. The primary (main) controller $G_c(z^{-1})$ can be designed by different approaches (for example digital PID control or methods based on algebraic approach). The outward feedback-loop through the block $G_d(z^{-1})$ in Fig. 1 is used to compensate for load disturbances and modelling errors.

Number of higher order industrial processes can be approximated by a reduced order model with a pure time-delay. In this paper the following second-order linear model with a time-delay is considered

$$G(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} z^{-d} = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} z^{-d} \quad (1)$$

The term z^{-d} represents the pure discrete time-delay. The time-delay is equal to dT_0 where T_0 is the sampling period.

Design of Polynomial 2DOF Controller

Previous simulation experiments demonstrated that polynomial theory is suitable method for design of the digital Smith Predictor. Polynomial control theory is based on the apparatus and methods of linear algebra

(see e.g. Kučera 1993). The polynomial Smith Predictor based on the digital pole assignment was designed in (Bobál et al. 2011). The design of the controller algorithm is based on the general block scheme of a closed-loop with two degrees of freedom (2DOF) according to Fig. 2.

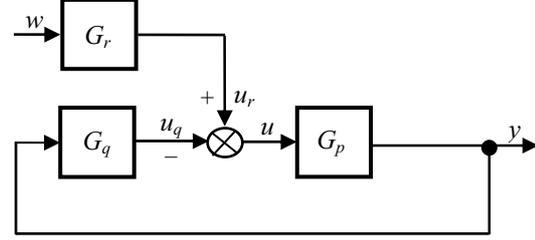


Figure 2: Block diagram of a closed loop 2DOF control system

The controlled process is given by the transfer function in the form

$$G_p(z^{-1}) = \frac{Y(z)}{U(z)} = \frac{B(z^{-1})}{A(z^{-1})} \quad (2)$$

where A and B are the second order polynomials. The controller contains the feedback part G_q and the feedforward part G_r . Then the digital controllers can be expressed in the form of discrete transfer functions

$$G_r(z^{-1}) = \frac{R(z^{-1})}{P(z^{-1})} = \frac{r_0}{(1 + p_1 z^{-1})(1 - z^{-1})} \quad (3)$$

$$G_q(z^{-1}) = \frac{Q(z^{-1})}{P(z^{-1})} = \frac{q_0 + q_1 z^{-1} + q_2 z^{-2}}{(1 + p_1 z^{-1})(1 - z^{-1})} \quad (4)$$

According to the scheme presented in Fig. 2 and equations (1) – (4) it is possible to derive the characteristic polynomial

$$A(z^{-1})P(z^{-1}) + B(z^{-1})Q(z^{-1}) = D_4(z^{-1}) \quad (5)$$

where

$$D_4(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + d_3 z^{-3} + d_4 z^{-4} \quad (6)$$

is the fourth degree characteristic polynomial.

The procedure leading to determination of polynomials Q , R and P in (3) and (4) can be briefly described as follows (see Bobál et al. 2005). A feedback part of the controller is given by a solution of the polynomial Diophantine equation (5). An asymptotic tracking is provided by a feedforward part of the controller given by a solution of the polynomial Diophantine equation

$$S(z^{-1})D_w(z^{-1}) + B(z^{-1})R(z^{-1}) = D_4(z^{-1}) \quad (7)$$

For a step-changing reference signal value, polynomial $D_w(z^{-1}) = 1 - z^{-1}$ and S is an auxiliary polynomial which does not enter into controller design.

A feedback controller to control a second-order system without time-delay will be derived from equation (5). A system of linear equations can be obtained using the uncertain coefficients method

$$\begin{bmatrix} b_1 & 0 & 0 & 1 \\ b_2 & b_1 & 0 & a_1 - 1 \\ 0 & b_2 & b_1 & a_2 - a_1 \\ 0 & 0 & b_2 & -a_2 \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ p_1 \end{bmatrix} = \begin{bmatrix} d_1 + 1 - a_1 \\ d_2 + a_1 - a_2 \\ d_3 + a_2 \\ d_4 \end{bmatrix} \quad (8)$$

For a step-changing reference signal value it is possible to derive the polynomial R from equation (7) by substituting $z = 1$

$$R = r_0 = \frac{D(1)}{B(1)} = \frac{1 + d_1 + d_2 + d_3 + d_4}{b_1 + b_2} \quad (9)$$

The 2DOF controller output is then given by

$$u(k) = r_0 w(k) - q_0 y(k) - q_1 y(k-1) - q_2 y(k-2) + (1 + p_1)u(k-1) + p_1 u(k-2) \quad (10)$$

Minimization of LQ Criterion

The linear quadratic methods try to minimize the quadratic criterion with penalization of the controller output

$$J = \sum_{k=0}^{\infty} \left\{ [w(k) - y(k)]^2 + q_u [u(k)]^2 \right\} \quad (11)$$

where q_u is the so-called penalization constant, which gives the rate of the controller output on the value of the criterion (where the constant at the first element of the criterion is considered equal to one). The standard procedure of the minimization of the criterion (11) is based on the state description of the system and leads to the solution of the Riccati Equation. In this paper, criterion minimization is realized through spectral factorization for the input-output description of the system (Bobál et al. 2005).

If the sequences of the values of both tracking error and input signal are considered as polynomials, it is possible to rewrite the criterion (11) using notation $\langle x(z) \rangle = x(0)$

$$J = \langle E(z)E(z^{-1}) + q_u U(z)U(z^{-1}) \rangle \quad (12)$$

where $E(z)$ and $U(z)$ are the conjugated polynomials to the polynomials $E(z^{-1})$ and $U(z^{-1})$, which means their negative powers are replaced by positive ones.

The tracking error polynomial

$$E(z^{-1}) = W(z^{-1}) - Y(z^{-1}) = \left[1 - \frac{B(z^{-1})R(z^{-1})}{A(z^{-1})P(z^{-1}) + B(z^{-1})Q(z^{-1})} \right] W(z^{-1}) \quad (13)$$

and the input signal polynomial

$$U(z^{-1}) = \frac{A(z^{-1})R(z^{-1})}{A(z^{-1})P(z^{-1}) + B(z^{-1})Q(z^{-1})} W(z^{-1}) \quad (14)$$

are substituted into criterion (11). It can be verified (Šebek and Kučera 1982) that the criterion is minimal if equation (5) is valid. The polynomial $D(z^{-1})$ is the result of spectral factorization according to the equation

$$A(z^{-1})q_u A(z) + B(z^{-1})B(z) = D(z^{-1})\delta D(z) \quad (15)$$

where δ is a constant chosen so that $d_0 = 1$. The spectral factorization of a polynomial leaves its stable part unchanged, while the unstable parts change to reciprocal ones (stable). Spectral factorization of polynomial of the first and the second degree could be computed simply; the procedure for the higher degrees must be performed iteratively.

While performing the spectral factorization of a polynomial of the second degree

$$M_2(z^{-1}) = m_0 + m_1 z^{-1} + m_2 z^{-2}$$

the following equation is solved

$$M_2(z^{-1})M_2(z) = D_2(z^{-1})\delta D_2(z) \quad (16)$$

where

$$D_2(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} \quad (17)$$

The products of the polynomials could be extended as

$$m_0 + m_1(z + z^{-1}) + m_2(z^2 + z^{-2}) = \delta(1 + d_1^2 + d_2^2) + \delta d_1(1 + d_2)(z + z^{-1}) + \delta d_2(z^2 + z^{-2}) \quad (18)$$

where the constants of the factorized polynomial on the left side of the equation (13) are combined into the coefficients m_0 , m_1 and m_2 . Comparing the left and the right side of equation (13), one obtains

$$m_0 = \delta(1 + d_1^2 + d_2^2); \quad m_1 = \delta d_1(1 + d_2); \quad m_2 = \delta d_2 \quad (19)$$

Solving equations (19), the following expressions are derived

$$\delta = \frac{\lambda + \sqrt{\lambda^2 - 4m_2^2}}{2}; \quad \lambda = \frac{m_0}{2} - m_2 + \sqrt{\left(\frac{m_0}{2} + m_2\right)^2 - m_1^2};$$

$$d_1 = \frac{m_1}{\delta + m_2}; \quad d_2 = \frac{m_2}{\delta} \quad (20)$$

Solving the spectral factorization of equation (15), an identical expression can be used, but is necessary to convert the left side of this equation to the form used in equation (16), thus

$$m_0 = q_u(1 + a_1^2 + a_2^2) + b_1^2 + b_2^2$$

$$m_1 = q_u(a_1 + a_1 a_2) + b_1 b_2; \quad m_2 = q_u a_2 \quad (21)$$

PRIMARY LQ CONTROLLER OF DIGITAL SMITH PREDICTOR

From the previous paragraph, it is obvious that using analytical spectral factorization, only two parameters of the second degree polynomial $D_2(z^{-1})$ (17) can be computed. This approach is applicable only for control of processes without time-delay (out of Smith Predictor). The primary controller in the digital Smith Predictor structure requires using the fourth degree polynomial $D_4(z^{-1})$ (6) in equations (5) and (7). From expression (17) it is obvious that polynomial

$$D_2(z) = z^2 + d_1z + d_2 \quad (22)$$

have two different real poles α, β or one complex conjugated pole $z_{1,2} = \alpha \pm j\beta$ (in the case of oscillatory systems). These poles must be included into polynomial

$$D_4(z) = z^4 + d_1z^3 + d_2z^2 + d_3z + d_4 \quad (23)$$

For both two types of the processes the suitable pole assignment was designed:

1st possibility:

Polynomial (18) has two different real poles α, β (computed from (17)) and user-defined real poles γ, δ . Then it is possible to write polynomial (18) as a product root of factor

$$D_4(z) = (z - \alpha)(z - \beta)(z - \gamma)(z - \delta) \quad (24)$$

and it is possible to express its individual parameters as

$$\begin{aligned} d_1 &= -(\alpha + \beta + \gamma + \delta) \\ d_2 &= \alpha\beta + \gamma\delta + (\alpha + \beta)(\gamma + \delta) \\ d_3 &= -[(\alpha + \beta)\gamma\delta + (\gamma + \delta)\alpha\beta] \\ d_4 &= \alpha\beta\gamma\delta \end{aligned} \quad (25)$$

2nd possibility:

Polynomial (18) has the complex conjugate pole $z_{1,2} = \alpha \pm j\beta$ (computed from (17)) and user-defined real poles γ, δ . Then polynomial (18) has the form

$$D_4(z) = (z - \alpha - j\beta)(z - \alpha + j\beta)(z - \gamma)(z - \delta) \quad (26)$$

and its individual parameters can be expressed as

$$\begin{aligned} d_1 &= -(2\alpha + \gamma + \delta) \\ d_2 &= 2\alpha(\gamma + \delta) + \alpha^2 + \beta^2 + \gamma\delta \\ d_3 &= -[2\alpha\gamma\delta + (\alpha^2 + \beta^2)(\gamma + \delta)] \\ d_4 &= (\alpha^2 + \beta^2)\gamma\delta \end{aligned} \quad (27)$$

The control algorithm based on the LQ control method contains the following steps:

1. The parameters of the polynomial $M_2(z^{-1})$ are computed according to equations (21).

2. The parameters of the polynomial $D_2(z^{-1})$ are computed according to equations (20).
3. If the polynomial (22) has the real poles α, β , its parameters are computed according to equations (25), otherwise, they are computed according to equations (27).
4. The controller parameters are computed using matrix equation (8) and equation (9).
5. The controller output is given by equation (10). Penalization of the controller output is performed by setting $q_u \geq 0$. With increased penalization constant, the amplitude of the controller output decreases and thereby, the flow of the process output is smoothed and any possible oscillations or instability are damped.

SIMULATION VERIFICATION AND RESULTS

A simulation verification of the designed predictive algorithm was performed in MATLAB/SIMULINK environment. A typical control scheme, which was used, is depicted in Fig. 3. This scheme is used for systems with time-delay of two sample steps. Individual blocks of the Simulink scheme correspond to blocks of the general control scheme presented in Fig. 2. It is possible to influence the output of the process with the non-measurable disturbance d .

The above mentioned Smith Predictor has universal usage for control of great deal of processes with time-delay. Therefore, four types of processes were chosen for simulation verification of controller algorithm. Consider the following continuous-time transfer functions:

- 1) Stable non-oscillatory $G_1(s) = \frac{2}{(s+1)(4s+1)} e^{-4s}$
- 2) Stable oscillatory $G_2(s) = \frac{2}{4s^2 + 2s + 1} e^{-4s}$
- 3) Non-minimum phase $G_3(s) = \frac{2(1-5s)}{(s+1)(4s+1)} e^{-4s}$
- 4) Unstable $G_4(s) = \frac{2(s+1)}{(2s-1)(4s+1)} e^{-4s}$

Let us now discretize them with a sampling period $T_0 = 2$ s, then the discrete forms are

$$G_1(z^{-1}) = \frac{0.4728z^{-1} + 0.2076z^{-2}}{1 - 0.7419z^{-1} + 0.0821z^{-2}} z^{-2}$$

$$G_2(z^{-1}) = \frac{0.6806z^{-1} + 0.4834z^{-2}}{1 - 0.7859z^{-1} + 0.3679z^{-2}} z^{-2}$$

$$G_3(z^{-1}) = \frac{-1.0978z^{-1} + 1.7783z^{-2}}{1 - 0.7419z^{-1} + 0.0821z^{-2}} z^{-2}$$

$$G_4(z^{-1}) = \frac{1.3248z^{-1} + 0.0274z^{-2}}{1 - 3.3248z^{-1} + 1.6487z^{-2}} z^{-2}$$

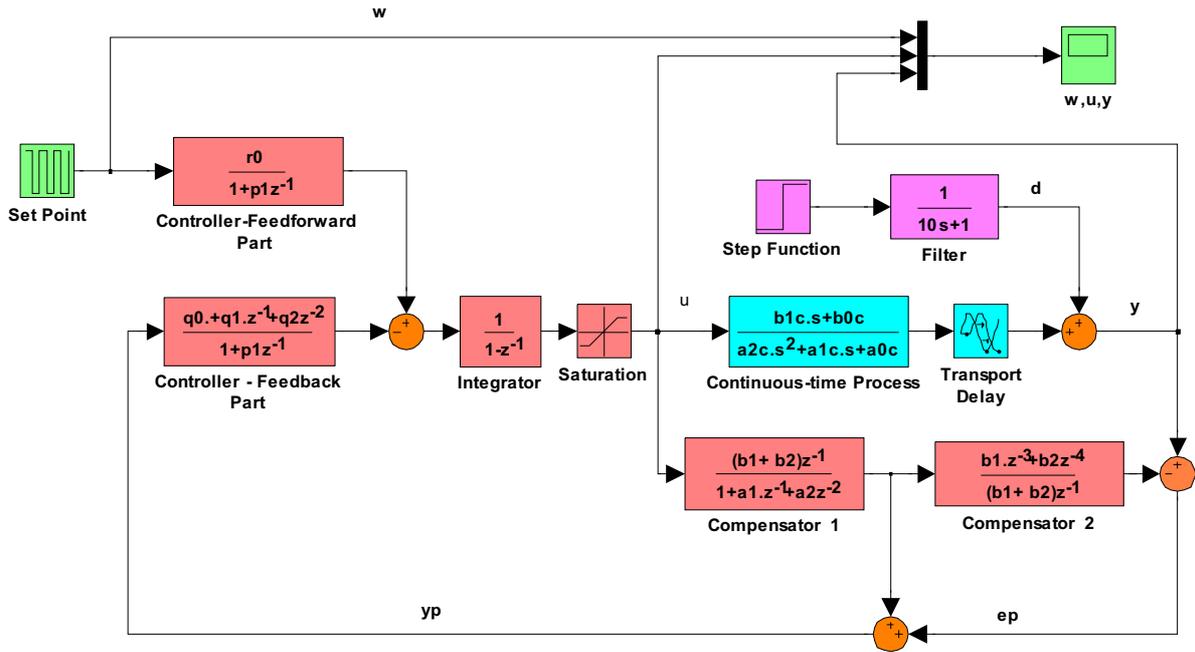


Figure 3: Simulink control scheme

The step responses of individual models are shown in Figs. 4 – 7.

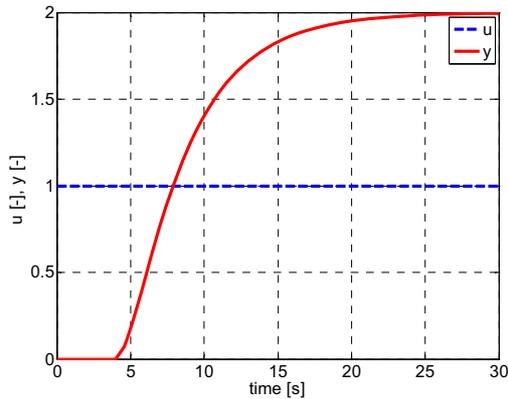


Figure 4: Step response of the model $G_1(s)$

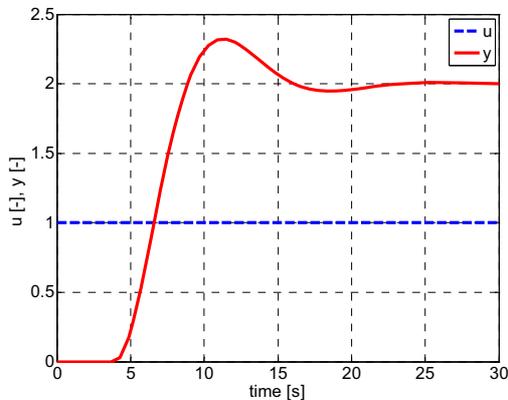


Figure 5: Step response of the model $G_2(s)$

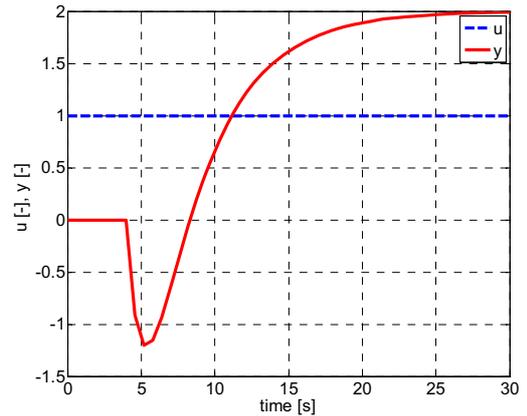


Figure 6: Step response of the model $G_3(s)$

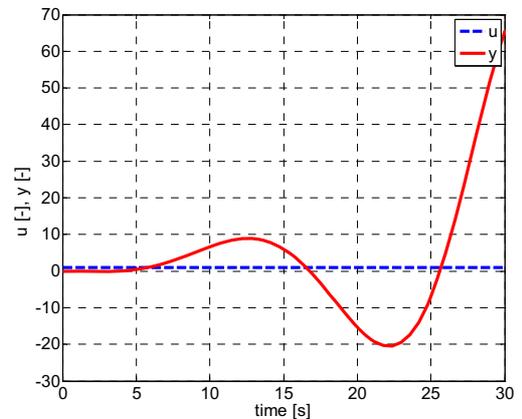


Figure 7: Step response of the model $G_4(s)$

The processes which are described by the above mentioned transfer functions were used in the Simulink control scheme for the verification of the dynamical

behaviour of individual closed control loops. In time 500 – 800 s an exponential external disturbance

$$d(t) = 0.25(1 - e^{-0.1t})$$

acted on the system output. The computed poles α, β and user-defined real poles γ, δ are introduced for individual simulation experiments including characteristic polynomial (25). For all experiments, the penalization factor was chosen $q_u = 1$.

Simulation control of model $G_1(z^{-1})$

The poles: $\alpha, \beta = 0.2130 \pm 0.2762i$; $\gamma = 0.1$; $\delta = 0.5$

The characteristic polynomial:

$$D_4(z) = z^4 - 1.0380z^3 + 0.3666z^2 - 0.0542z + 0.0027$$

The courses of the control variables are shown in Fig. 8, the quality of control is very good.

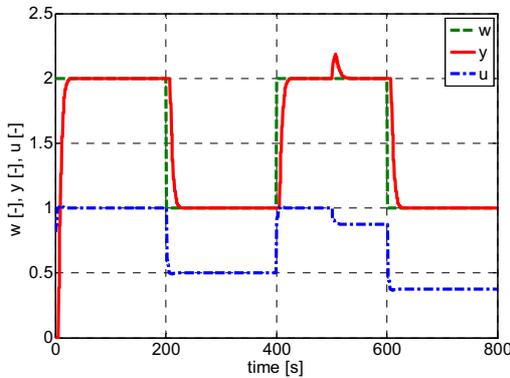


Figure 8: Control of the model $G_1(z^{-1})$

Simulation control of model $G_2(z^{-1})$

The poles: $\alpha, \beta = 0.1451 \pm 0.3820i$; $\gamma = 0.1$; $\delta = 0.5$

The characteristic polynomial:

$$D_4(z) = z^4 - 0.8902z^3 + 0.3911z^2 - 0.1147z + 0.0027$$

The courses of the control variables are shown in Fig. 9, the quality of control is very good.

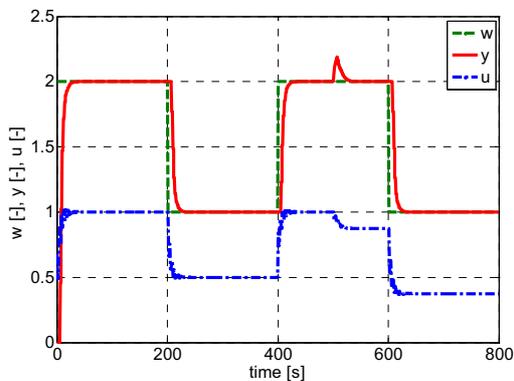


Figure 9: Control of the model $G_2(z^{-1})$

Simulation control of model $G_3(z^{-1})$

The poles: $\alpha = 0.6153$; $\beta = 0.0325$; $\gamma = 0.1$; $\delta = 0.75$

The characteristic polynomial:

$$D_4(z) = z^4 - 1.4973z^3 + 0.6449z^2 - 0.0653z + 0.0015$$

The courses of the control variables are shown in Fig. 9. The process output y has typical character for the control of the non-minimum phase system (undershoot of y in the initial time-interval). The stability of a control-loop is very dependent on pole δ . For small δ the control loop is unstable, for suitable chosen δ , the quality of control is very good.

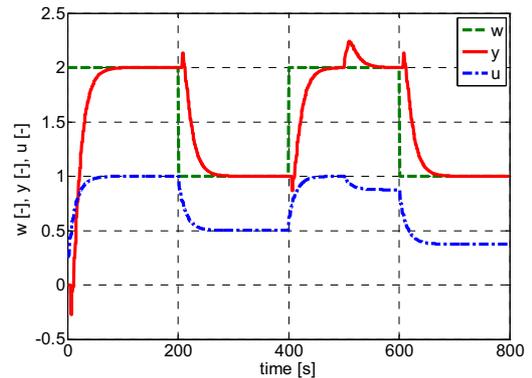


Figure 10: Control of model $G_3(z^{-1})$

Simulation control of model $G_4(z^{-1})$

The poles: $\alpha, \beta = 0.3470 \pm 0.1720i$; $\gamma = 0.1$; $\delta = 0.5$

The characteristic polynomial:

$$D_4(z) = z^4 - 1.2940z^3 + 0.6164z^2 - 0.1247z + 0.0075$$

The courses of the control variables are shown in Fig. 10, the quality of control is very good.

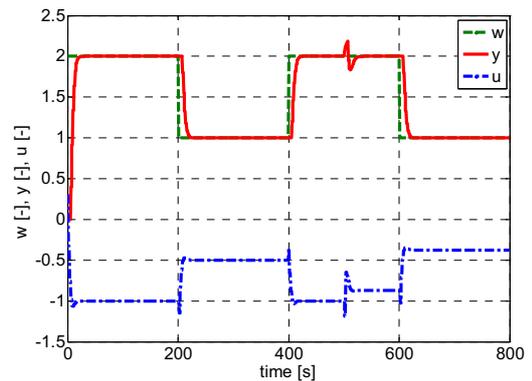


Figure 10: Control of the model $G_4(z^{-1})$

CONCLUSION

The contribution presents new generalized strategy for design of the polynomial digital Smith Predictor for control systems with time-delay. The primary

controller is based on minimization of the linear quadratic criterion. Minimization of criterion is realized through spectral factorization. This controller was derived purposely by analytical way (without utilization of numerical methods) to obtain algorithms with easy implementability in industrial practice. Four models of control processes were used for simulation verification. Main contribution of the proposed method is the universal applicability of digital Smith Predictor for unstable processes. The designed predictive controller was successfully verified not only by simulation but also in real-time laboratory conditions for control of a heat exchanger.

ACKNOWLEDGEMENTS

This article was created with support of Operational Programme Research and Development for Innovations co-funded by the European Regional Development Fund (ERDF), national budget of Czech Republic within the framework of the Centre of Polymer Systems project (reg. number: CZ.1.05/2.1.00/03.0111).

REFERENCES

- Bobál, V., Böhm, J., Fessl, J. and J. Macháček. 2005. *Digital Self-tuning Controllers: Algorithms, Implementation and Applications*. Springer-Verlag, London.
- Bobál, V., Chalupa, P., Dostál, P. and M. Kubalčík. 2011. "Design and simulation verification of self-tuning Smith predictors." *International Journal of Mathematics and Computers in Simulation* 5, 342-351.
- Bobál, V., Chalupa, P., Novák, J. and P. Dostál. 2012a. "MATLAB Toolbox for CAD of self-tuning of time-delay processes." In *Proc. of the International Workshop on Applied Modelling and Simulation*, Roma, 44 – 49.
- Bobál, V., Chalupa, P. and J. Novák. 2012b. "Toolbox for CAD and Verification of Digital Adaptive Control Time-Delay Systems." Available from http://nod32.fai.utb.cz/promotion/Software_OBD/Time_Delay_Tool.zip.
- Camacho, E. F. and C. Bordons. 2004. *Model Predictive Control*. Springer-Verlag, London.
- Hang, C. C., Lim, K. W. and B. W. Chong. 1989. "A dual-rate digital Smith predictor." *Automatica* 20, 1-16.
- Hang, C. C., Tong, H. L. and K. H. Weng. 1993. *Adaptive Control*, North Carolina: Instrument Society of America.
- Kučera, V. (1993). "Diophantine equations in control – a survey." *Automatica* 29, 1361-1375.
- Normey-Rico, J. E. and E. F. Camacho. 1998. "Dead-time compensators: A unified approach. In *Proceedings of IFAC Workshop on Linear Time Delay Systems (LDTS'98)*, Grenoble, France, 141-146.
- Normey-Rico, J. E. and E. F. Camacho. 2007. *Control of Dead-time Processes*, Springer-Verlag, London.
- Palmor, Z. J. and Y. Halevi. 1990. "Robustness properties of sampled-data systems with dead time compensators." *Automatica* 26, 637-640.
- Smith, O. J. 1957. "Closed control of loops." *Chem. Eng. Progress*, 53, 217-219.
- Šebek, M. and V. Kučera. 1982. "Polynomial approach to quadratic tracking in discrete linear systems." *IEEE Trans. Automatic. Control* AC-27, 1248-1250.
- Vogel, E. F. and T. F. Edgar (1980). "A new dead time compensator for digital control." In *Proceedings ISA Annual Conference*, Houston, USA, 29-46.

AUTHOR BIOGRAPHIES



VLADIMÍR BOBÁL graduated in 1966 from the Brno University of Technology, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are adaptive and predictive control, system identification and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.



MAREK KUBALČÍK graduated in 1993 from the Brno University of Technology in Automation and Process Control. He received his Ph.D. degree in Technical Cybernetics at Brno University of Technology in 2000. From 1993 to 2007 he worked as senior lecturer at the Faculty of Technology, Brno University of Technology. From 2007 he has been working as an associate professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. Current work covers following areas: control of multivariable systems, self-tuning controllers, predictive control. His e-mail address is: kubalcik@fai.utb.cz.



PETR DOSTÁL studied at the Technical University of Pardubice, Czech Republic, where he obtained his master degree in 1968 and Ph.D. degree in Technical Cybernetics in 1979. In the year 2000 he became professor in Process Control. He is now head of the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interests are modelling and simulation of continuous-time chemical processes, polynomial methods, optimal and adaptive control. You can contact him on email address dostalp@fai.utb.cz.



STANISLAV TALAŠ studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2013. He now attends PhD. study in the Department of Process Control. His e-mail address is talas@fai.utb.cz.

SIMULATION OF MULTIVARIABLE CONTINUOUS-TIME DECOUPLING CONTROL

Marek Kubalčík and Vladimír Bobál
Tomas Bata University in Zlín
Faculty of Applied Informatics
Nad Stráněmi 4511, 760 05, Zlín, Czech Republic
E-mail: kubalcik@fai.utb.cz

KEYWORDS

Multivariable control, Control algorithms, Adaptive control, Polynomial methods, Simulation.

ABSTRACT

The paper is focused on an implementation of a decoupling multivariable controller in the Matlab/Simulink environment. The control algorithm is based on polynomial theory and pole – placement. A decoupling compensator is used to suppress interactions between control loops. The controller was realized both with fixed parameters and with recursive identification of a model of the controlled system. The internal structure of the controller enables its easy modification and implementation of further similar control algorithms.

INTRODUCTION

Typical technological processes require the simultaneous control of several variables related to one system. Each input may influence all system outputs. The design of a controller for such a system must be quite sophisticated if the system is to be controlled adequately. There are many different methods of controlling MIMO (multi input – multi output) systems. Several of these use decentralized PID controllers (Cui and Jacobsen, 2002) others apply single input-single-output (SISO) methods extended to cover multiple inputs (Chien et al, 1987). The classical approach to the control of multi-input–multi-output (MIMO) systems is based on the design of a matrix controller to control all system outputs at one time. The basic advantage of this approach is its ability to achieve optimal control performance because the controller can use all the available information about the controlled system. Controllers are based on various approaches and various mathematical models of controlled processes. A standard technique for MIMO control systems uses polynomial methods (Kučera, 1980, Kučera 1991, Vidyasagar 1985) and is also used in this paper. Controller synthesis is reduced to the solution of linear Diophantine equations (Kučera, 1993) . One controller, which enables decoupling control of TITO (two input-two output) systems, is presented. The proposed control algorithm applies a decoupling compensator (Krishnawamy et al, 1991, Peng, 1990, Tade et al, 1986) to suppress undesired interactions

between control loops. The controller was realized both with fixed parameters and with recursive identification of a model of the controlled system.

For purposes of simulation, the controller was realized in the Matlab/Simulink environment as a mask of subsystem. It can be easily inserted into Simulink schemes of the closed loop. No initialization is needed before the simulation start. Blocks for computation of the control law and for recursive identification were realized as S-functions. The internal structure of the controller enables its easy modification and implementation of further similar control algorithms. A simulation experiment is also introduced.

MODEL OF THE CONTROLLED SYSTEM

A general transfer matrix of a two-input–two-output system with significant cross-coupling between the control loops is expressed as

$$\mathbf{G}(s) = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} \quad (1)$$

$$\mathbf{Y}(s) = \mathbf{G}(s)\mathbf{U}(s) \quad (2)$$

where $\mathbf{U}(s)$ and $\mathbf{Y}(s)$ are vectors of the manipulated variables and the controlled variables.

$$\mathbf{Y}(s) = [y_1(s), y_2(s)]^T \quad \mathbf{U}(s) = [u_1(s), u_2(s)]^T \quad (3)$$

It may be assumed that the transfer matrix can be transcribed to the following form of the matrix fraction:

$$\mathbf{G}(s) = \mathbf{A}^{-1}(s)\mathbf{B}(s) = \mathbf{B}_1(s)\mathbf{A}_1^{-1}(s) \quad (4)$$

where the polynomial matrices $\mathbf{A} \in R_{22}[s]$, $\mathbf{B} \in R_{22}[s]$ represent the left coprime factorization of matrix $\mathbf{G}(s)$ and the matrices $\mathbf{A}_1 \in R_{22}[s]$, $\mathbf{B}_1 \in R_{22}[s]$ represent the right coprime factorization of $\mathbf{G}(s)$. The further described algorithm is based on a model with polynomials of second order. This model proved to be effective for control of several TITO laboratory processes (Kubalčík and Bobál, 2006), where controllers based on a model with polynomials of the first order failed. In case of decoupling control using a compensator it is useful to consider matrix $\mathbf{A}(s)$ as diagonal. The reason is explained in the following section.

$$\mathbf{A}(s) = \begin{bmatrix} s^2 + a_1s + a_2 & 0 \\ 0 & s^2 + a_7s + a_8 \end{bmatrix} \quad (5)$$

$$\mathbf{B}(s) = \begin{bmatrix} b_1s + b_2 & b_3s + b_4 \\ b_5s + b_6 & b_7s + b_8 \end{bmatrix} \quad (6)$$

Differential equations describing dynamical behavior of the system are as follows

$$y_1'' + a_1y_1' + a_2y_1 = b_1u_1' + b_2u_1 + b_3u_2' + b_4u_2 \quad (7)$$

$$y_2'' + a_3y_2' + a_4y_2 = b_5u_1' + b_6u_1 + b_7u_2' + b_8u_2 \quad (8)$$

DESIGN OF THE DECOUPLING CONTROLLER

One of possible approaches to control of multivariable systems is the serial insertion of a compensator ahead of the system (Krishnawamy et al, 1991, Peng, 1990, Tade et al, 1986). The compensator then becomes a part of the controller. The objective, in this case, is to suppress undesirable interactions between the input and output variables so that each input affects only one controlled variable. The block diagram for this kind of system is shown in Figure 1 (\mathbf{R} is a transfer matrix of a controller and \mathbf{C} is a decoupling compensator).

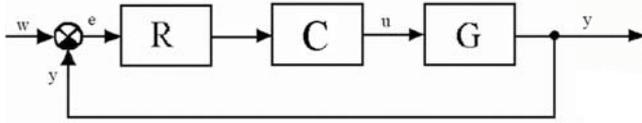


Figure 1: Closed loop system with compensator

The resulting transfer function \mathbf{H} (the operator s will be omitted from some operations for the purpose of simplification) is then determined by

$$\mathbf{H} = \mathbf{GC} = \mathbf{A}^{-1}\mathbf{BC} = \mathbf{A}^{-1}\mathbf{H}_1 \quad (9)$$

The decoupling conditions are fulfilled when matrix \mathbf{H} is diagonal. As it was mentioned above the matrix \mathbf{A} is supposed to be diagonal. The reason for this simplification is apparent from equation (9). When matrix \mathbf{A} is assumed to be non-diagonal it has to be included into the compensator in order to obtain a diagonal matrix \mathbf{H} . The order of the controller and consequently complexity of its design would increase. The compensator is defined as the adjugated matrix \mathbf{B}

$$\mathbf{C} = \text{adj}(\mathbf{B}) \quad (10)$$

The matrix \mathbf{H}_1 then takes following form

$$\mathbf{H}_1 = \mathbf{B}\text{adj}(\mathbf{B}) = \begin{bmatrix} \det(\mathbf{B}) & 0 \\ 0 & \det(\mathbf{B}) \end{bmatrix} \quad (11)$$

Generally, the vector of input reference signals \mathbf{W} is specified as

$$\mathbf{W}(s) = \mathbf{F}_w^{-1}(s)\mathbf{h}(s) \quad (12)$$

Further the reference signals are considered as step functions. In this case \mathbf{h} is a vector of constants and \mathbf{F}_w is expressed as

$$\mathbf{F}_w(s) = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \quad (13)$$

The controller can be described both by left and right matrix fractions as well as the controlled system

$$\mathbf{G}_R(s) = \mathbf{P}^{-1}(s)\mathbf{Q}(s) = \mathbf{Q}_1(s)\mathbf{P}_1^{-1}(s) \quad (14)$$

In order to achieve asymptotic tracking of the reference signal, an integrator must be incorporated into the controller. The controller including the integrator can be defined as

$$\mathbf{R} = \mathbf{F}^{-1}\mathbf{Q}_1\mathbf{P}_1^{-1} \quad (15)$$

The component \mathbf{F} is the integrator. The resulting matrix of the controller can be then defined as follows

$$\mathbf{CR} = \mathbf{CF}^{-1}\mathbf{Q}_1\mathbf{P}_1^{-1} \quad (16)$$

It is possible to derive an equation for the system output, which can be modified by matrix operations to the form

$$\mathbf{Y} = \mathbf{P}_1(\mathbf{AFP}_1 + \mathbf{H}_1\mathbf{Q}_1)^{-1}\mathbf{H}_1\mathbf{Q}_1\mathbf{P}_1\mathbf{W} \quad (17)$$

The determinant of the matrix in the denominator ($\mathbf{AFP}_1 + \mathbf{H}_1\mathbf{Q}_1$) is the characteristic polynomial of the MIMO system. The roots of this polynomial matrix determine the behaviour of the closed loop system. They must be placed on the left side of the Gauss complex plane for the system to be stable. Conditions of BIBO stability can be defined by the following Diophantine matrix equation:

$$\mathbf{AFP}_1 + \mathbf{H}_1\mathbf{Q}_1 = \mathbf{M} \quad (18)$$

where $\mathbf{M} \in R_{22}[s]$ is a stable diagonal polynomial matrix. If the system has the same number of inputs and outputs, matrix \mathbf{M} can be chosen as diagonal, which allows easier computation of the controller parameters. Correct pole placement of the matrix \mathbf{M} is very important for good control performance.

$$\mathbf{M}(s) = \begin{bmatrix} s^5 + m_1s^4 + & 0 \\ +m_2s^3 + m_3s^2 + m_4s + m_5 & \\ 0 & s^5 + m_6s^4 + m_7s^3 + \\ & + m_8s^2 + m_9s + m_{10} \end{bmatrix} \quad (19)$$

The degree of the controller polynomial matrices depends on the internal properness of the closed loop. The structures of matrices \mathbf{P}_1 and \mathbf{Q}_1 were chosen so that the number of unknown controller parameters equals the number of algebraic equations resulting from the solution of the Diophantine equation (18) using the method of uncertain coefficients:

$$\mathbf{P}_1(s) = \begin{bmatrix} s^2 + p_1s + p_2 & 0 \\ 0 & s^2 + p_3s + p_4 \end{bmatrix} \quad (20)$$

$$\mathbf{Q}_1(s) = \begin{bmatrix} q_1 s^2 + q_2 s + q_3 & 0 \\ 0 & q_4 s^2 + q_5 s + q_6 \end{bmatrix} \quad (21)$$

For simplification, it was computed the determinant $\det(\mathbf{B})$:

$$\det(\mathbf{B}) = (b_1 b_7 - b_3 b_5) s^2 + (b_1 b_8 + b_2 b_7 - b_4 b_5 - b_3 b_6) s + (b_2 b_8 - b_4 b_6) = db_3 s^2 + db_2 s + db_1 \quad (22)$$

The solution of the Diophantine equation results in a set of 10 algebraic equations with unknown controller parameters. Using matrix notation, the algebraic equations are expressed in the following form.

$$\begin{bmatrix} 1 & 0 & db_3 & 0 & 0 \\ a_1 & 1 & db_2 & db_3 & 0 \\ a_2 & a_1 & db_1 & db_2 & db_3 \\ 0 & a_2 & 0 & db_1 & db_2 \\ 0 & 0 & 0 & 0 & db_1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} m_1 - a_1 \\ m_2 - a_2 \\ m_3 \\ m_4 \\ m_5 \end{bmatrix} \quad (23)$$

$$\begin{bmatrix} 1 & 0 & db_3 & 0 & 0 \\ a_3 & 1 & db_2 & db_3 & 0 \\ a_4 & a_3 & db_1 & db_2 & db_3 \\ 0 & a_4 & 0 & db_1 & db_2 \\ 0 & 0 & 0 & 0 & db_1 \end{bmatrix} \begin{bmatrix} p_3 \\ p_4 \\ q_4 \\ q_5 \\ q_6 \end{bmatrix} = \begin{bmatrix} m_6 - a_3 \\ m_7 - a_4 \\ m_8 \\ m_9 \\ m_{10} \end{bmatrix} \quad (24)$$

The control law is defined as:

$$\mathbf{FU} = \text{adj}(\mathbf{B})\mathbf{Q}_1\mathbf{P}_1^{-1}\mathbf{E} \quad (25)$$

where \mathbf{E} is a vector of control errors. This matrix equation can be transcribed to the differential equations of the controller

$$\begin{aligned} & u_1^{(5)} + u_1^{(4)}(p_3 + p_1) + u_1^{(3)}(p_4 + p_1 p_3 + p_2) + u_1^{(2)}(p_1 p_4 + p_2 p_3) + u_1^{(1)} p_2 p_4 = \\ & = e_1^{(5)} b_7 q_1 + e_1^{(4)}(b_7 q_2 + b_8 q_1 + p_3 b_7 q_1) + e_1^{(3)}(b_7 q_3 + b_8 q_2 + p_4 b_7 q_1 + p_3 b_7 q_2 + p_3 b_8 q_1) + \\ & + e_1^{(2)}(b_8 q_3 + p_3 b_7 q_3 + p_3 b_8 q_2 + p_4 b_7 q_2 + p_4 b_8 q_1) + e_1^{(1)}(p_4 b_7 q_3 + p_4 b_8 q_2 + p_3 b_8 q_3) + e_1 p_4 b_8 q_3 - \\ & - e_2^{(5)} b_3 q_4 - e_2^{(4)}(b_3 q_5 + b_4 q_4 + p_1 b_3 q_4) - e_2^{(3)}(b_3 q_6 + b_4 q_5 + p_1 b_3 q_5 + p_1 b_4 q_4 + p_3 b_3 q_4) - \\ & - e_2^{(2)}(b_4 q_6 + p_1 b_3 q_6 + p_1 b_4 q_5 + p_2 b_3 q_5 + p_2 b_4 q_4) - e_2^{(1)}(p_2 b_3 q_6 + p_2 b_4 q_5) - e_2 p_2 b_4 q_6 \end{aligned} \quad (26)$$

$$\begin{aligned} & u_2^{(5)} + u_2^{(4)}(p_3 + p_1) + u_2^{(3)}(p_4 + p_1 p_3 + p_2) + u_2^{(2)}(p_1 p_4 + p_2 p_3) + u_2^{(1)} p_2 p_4 = \\ & = -e_1^{(5)} b_3 q_1 - e_1^{(4)}(b_3 q_2 + b_4 q_1 + p_3 b_3 q_1) - e_1^{(3)}(b_3 q_3 + b_4 q_2 + p_4 b_3 q_1 + p_3 b_3 q_2 + p_3 b_4 q_1) - \\ & - e_1^{(2)}(b_4 q_3 + p_4 b_3 q_3 + p_3 b_3 q_3 + p_4 b_3 q_2 + p_3 b_4 q_2) - e_1^{(1)}(p_3 b_3 q_3 + p_4 b_3 q_3 + p_4 b_4 q_2) - e_1 p_4 b_3 q_3 + \\ & + e_2^{(5)} b_4 q_4 + e_2^{(4)}(b_4 q_5 + b_4 q_4 + p_1 b_4 q_4) + e_2^{(3)}(b_4 q_6 + b_4 q_5 + p_2 b_4 q_4 + p_1 b_4 q_5 + p_1 b_2 q_4) + \\ & + e_2^{(2)}(b_2 q_6 + p_2 b_4 q_5 + p_2 b_2 q_4 + p_1 b_4 q_6 + p_1 b_2 q_5) - e_2^{(1)}(p_2 b_4 q_6 + p_2 b_2 q_5 + p_1 b_2 q_6) + e_2 p_2 b_2 q_6 \end{aligned} \quad (27)$$

For purposes of simulation, the controller was realized in the Matlab/Simulink environment as an S-function. It was then necessary to obtain its state equations. Further there it is introduced a conversion of the first differential equation (26) to the state equations. The second differential equation (27) was conversed similarly. Equation (26) can be itemized as follows

$$\begin{aligned} & u_{1A}^{(5)} + u_{1A}^{(4)}(p_3 + p_1) + u_{1A}^{(3)}(p_4 + p_1 p_3 + p_2) + u_{1A}^{(2)}(p_1 p_4 + p_2 p_3) + u_{1A}^{(1)} p_2 p_4 = \\ & = e_1^{(5)} b_7 q_1 + e_1^{(4)}(b_7 q_2 + b_8 q_1 + p_3 b_7 q_1) + e_1^{(3)}(b_7 q_3 + b_8 q_2 + p_4 b_7 q_1 + p_3 b_7 q_2 + p_3 b_8 q_1) + \\ & + e_1^{(2)}(b_8 q_3 + p_3 b_7 q_3 + p_3 b_8 q_2 + p_4 b_7 q_2 + p_4 b_8 q_1) + e_1^{(1)}(p_4 b_7 q_3 + p_4 b_8 q_2 + p_3 b_8 q_3) + e_1 p_4 b_8 q_3 \end{aligned} \quad (28)$$

$$\begin{aligned} & u_{1B}^{(5)} + u_{1B}^{(4)}(p_3 + p_1) + u_{1B}^{(3)}(p_4 + p_1 p_3 + p_2) + u_{1B}^{(2)}(p_1 p_4 + p_2 p_3) + u_{1B}^{(1)} p_2 p_4 = \\ & = -e_2^{(5)} b_3 q_4 - e_2^{(4)}(b_3 q_5 + b_4 q_4 + p_1 b_3 q_4) - e_2^{(3)}(b_3 q_6 + b_4 q_5 + p_1 b_3 q_5 + p_1 b_4 q_4 + p_3 b_3 q_4) - \\ & - e_2^{(2)}(b_4 q_6 + p_1 b_3 q_6 + p_1 b_4 q_5 + p_2 b_3 q_5 + p_2 b_4 q_4) - e_2^{(1)}(p_2 b_3 q_6 + p_2 b_4 q_5) - e_2 p_2 b_4 q_6 \end{aligned} \quad (29)$$

Equation (28) can be transcribed to the transfer function. It is also possible to establish an auxiliary variable Z

$$\begin{aligned} G(s) &= \frac{b_7 q_1 s^5 + (b_7 q_2 + b_8 q_1 + p_3 b_7 q_1) s^4 + (p_4 b_7 q_3 + p_4 b_8 q_2 + p_3 b_8 q_3) s^3 +}{s^5 + (p_3 + p_1) s^4 + (p_4 + p_1 p_3 + p_2) s^3 + (p_1 p_4 + p_2 p_3) s^2 + p_2 p_4 s} \\ &+ \frac{(b_7 q_3 + b_8 q_2 + p_4 b_7 q_1 + p_3 b_7 q_2 + p_3 b_8 q_1) s^3 + p_4 b_8 q_3 +}{(b_8 q_3 + p_3 b_7 q_3 + p_3 b_8 q_2 + p_4 b_7 q_2 + p_4 b_8 q_1) s^2} \\ &= \frac{U_{1A}}{E_1} = \frac{U_{1A}}{Z} \frac{Z}{E_1} \end{aligned} \quad (30)$$

By means of the variable Z it is possible to define following equations

$$\begin{aligned} & b_7 q_1 z^{(5)} + (b_7 q_2 + b_8 q_1 + p_3 b_7 q_1) z^{(4)} + (b_7 q_3 + b_8 q_2 + p_4 b_7 q_1 + p_3 b_7 q_2 + p_3 b_8 q_1) z^{(3)} + \\ & + (b_8 q_3 + p_3 b_7 q_3 + p_3 b_8 q_2 + p_4 b_7 q_2 + p_4 b_8 q_1) z^{(2)} + (p_4 b_7 q_3 + p_4 b_8 q_2 + p_3 b_8 q_3) z^{(1)} + \\ & + p_4 b_8 q_3 z = u_{1A} \end{aligned} \quad (31)$$

$$z^{(5)} + (p_3 + p_1) z^{(4)} + (p_4 + p_1 p_3 + p_2) z^{(3)} + (p_1 p_4 + p_2 p_3) z^{(2)} + p_2 p_4 z^{(1)} = e_1 \quad (32)$$

Equation (32) can be converted to a set of differential equations of the first order (state equations). Choice of the state variables is as follows

$$x_1 = z \quad x_2 = z' \quad x_3 = z'' \quad x_4 = z''' \quad x_5 = z^{(4)} \quad (33)$$

And the state equations are

$$\begin{aligned} & x_1' = x_2 \\ & x_2' = x_3 \\ & x_3' = x_4 \\ & x_4' = x_5 \\ & x_5' = e_1 - (p_3 + p_1) x_5 - (p_4 + p_1 p_3 + p_2) x_4 - (p_1 p_4 + p_2 p_3) x_3 - p_2 p_4 x_2 \end{aligned} \quad (34)$$

On the basis of the state variables, which are substituted to equation (31), it is possible to derive the first part of the manipulated variable u_{1A}

$$\begin{aligned} u_{1A} &= b_7 q_1 (e_1 - (p_3 + p_1) x_5 + (p_4 + p_1 p_3 + p_2) x_4 + (p_1 p_4 + p_2 p_3) x_3 + p_2 p_4 x_2) + \\ &+ (b_7 q_2 + b_8 q_1 + p_3 b_7 q_1) x_5 + (b_7 q_3 + b_8 q_2 + p_4 b_7 q_1 + p_3 b_7 q_2 + p_3 b_8 q_1) x_4 + \\ &+ (b_8 q_3 + p_3 b_7 q_3 + p_3 b_8 q_2 + p_4 b_7 q_2 + p_4 b_8 q_1) x_3 + (p_4 b_7 q_3 + p_4 b_8 q_2 + p_3 b_8 q_3) x_2 + \\ &+ p_4 b_8 q_3 x_1 \end{aligned} \quad (35)$$

Similarly it is possible to transcribe equation (29)

$$\begin{aligned} & -b_3 q_4 z^{(5)} - (b_3 q_5 + b_4 q_4 + p_1 b_3 q_4) z^{(4)} - (b_3 q_6 + b_4 q_5 + p_1 b_3 q_5 + p_1 b_4 q_4 + p_3 b_3 q_4) z^{(3)} - \\ & - (b_4 q_6 + p_1 b_3 q_6 + p_1 b_4 q_5 + p_2 b_3 q_5 + p_2 b_4 q_4) z^{(2)} - (p_2 b_3 q_6 + p_2 b_4 q_5) z^{(1)} - \\ & - p_2 b_4 q_6 z = u_{1B} \end{aligned} \quad (36)$$

$$z^{(5)} + (p_3 + p_1) z^{(4)} + (p_4 + p_1 p_3 + p_2) z^{(3)} + (p_1 p_4 + p_2 p_3) z^{(2)} + p_2 p_4 z^{(1)} = e_2 \quad (37)$$

State variables were chosen similarly as in the previous case

$$x_6 = z \quad x_7 = z' \quad x_8 = z'' \quad x_9 = z''' \quad x_{10} = z^{(4)} \quad (38)$$

The state equations are then as follows

$$\begin{aligned} x_6' &= x_7 \\ x_7' &= x_8 \\ x_8' &= x_9 \\ x_9' &= x_{10} \\ x_{10}' &= e_2 - (p_3 + p_1)x_{10} - (p_4 + p_1p_3 + p_2)x_9 - (p_1p_4 + p_2p_3)x_8 - p_2p_4x_7 \end{aligned} \quad (39)$$

The second part of the manipulated variable u_{1B} can be computed similarly like the part u_{1A} by substitution of the state variables to equation (36)

$$\begin{aligned} u_{1B} &= -b_3q_4(e_2 - (p_3 + p_1)x_{10} + (p_4 + p_1p_3 + p_2)x_9 + (p_1p_4 + p_2p_3)x_8 + p_2p_4x_7) - \\ &\quad - (b_3q_5 + b_4q_4 + p_1b_3q_4)x_{10} - (b_3q_6 + b_4q_5 + p_1b_3q_5 + p_1b_4q_4 + p_3b_3q_4)x_9 - \\ &\quad - (b_4q_6 + p_1b_3q_6 + p_1b_4q_5 + p_2b_3q_5 + p_2b_4q_4)x_8 - (p_1b_4q_6 + p_2b_3q_6 + p_2b_4q_5)x_7 - \\ &\quad - p_2b_4q_6x_6 \end{aligned} \quad (40)$$

The manipulated variable u_1 is then defined by the following sum

$$u_1 = u_{1A} + u_{1B} \quad (41)$$

An expression for computation of the manipulated variable u_2 is obtained similarly on the basis of differential equation (27).

RECURSIVE IDENTIFICATION

The controller was also realized as a self-tuning controller with recursive identification of a model of the controlled system. The recursive least squares method (Bobál, 2005) proved to be effective for self-tuning controllers and was used as the basis for our algorithm. For our two-variable example we considered the disintegration of the identification into two independent parts. It is not possible to measure directly input and output derivatives of a system in case of continuous – time control loop. One of the possible approaches to this problem is establishing of filters and filtered variables to substitute the primary variables. This approach is described in detail in (Wahlberg, 1990). The filtered variables are then used in the recursive identification procedure.

IMPLEMENTATION OF THE CONTROLLER

A Simulink scheme of the closed loop system is in Figure 3. The controller uses two input signals and provides two outputs. The inputs are the reference signals (w_1, w_2) and the vector of actual outputs of controlled process (y_1, y_2). The main controller output is the vector of control signals – the input signals of the controlled process. The second controller output consists of the current parameter estimates of the controlled process model.

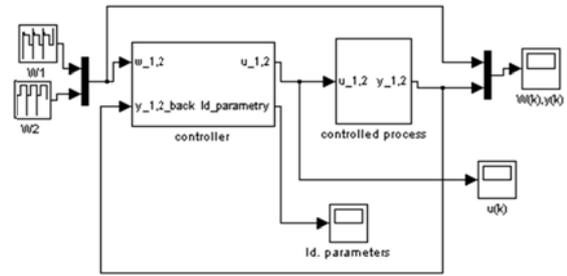


Figure 2: SIMULINK scheme of closed loop system

The controller is constructed as a mask of a subsystem, which consists of Simulink blocks and has inputs, outputs and parameters. Internal controller structure consists of Simulink blocks which provide, among others, the possibility of easy creation of a new controller by a modification of the proposed controller. The structure of the controller is presented in Figure 3. It consists of two basic parts: an on – line identification block and a block for computation of the control law and the manipulated variables. Controller's parameters are set using dialog windows.

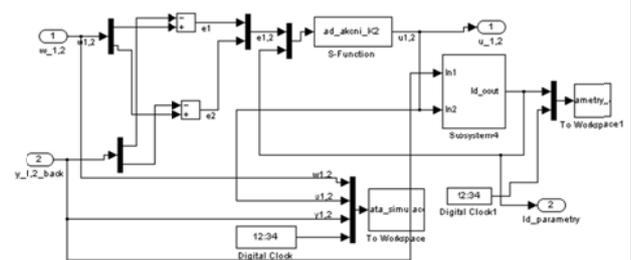


Figure 3: Internal structure of the controller

In Figure 4 is an internal structure of the subsystem for the system identification. It consists of the filters for filtering of the controlled and manipulated variables and a block for recursive identification.

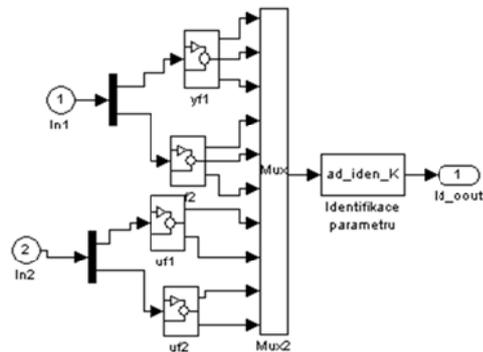


Figure 4: Filtering of variables and recursive identification

SIMULATION VERIFICATION

Verification by simulation was carried out on a range of plants with various dynamics. The control of the model below is given here as an example. The controller's

synthesis is based on the model with diagonal matrix A , which is obtained by recursive identification and which describes the dynamics of the system with full matrix A .

$$A(s) = \begin{bmatrix} s^2 + 2s + 0,7 & 0,2s + 0,4 \\ -0,5s - 0,1 & s^2 + 2s + 0,7 \end{bmatrix} \quad (42)$$

$$B(s) = \begin{bmatrix} 0,5s + 0,2 & 0,1s + 0,3 \\ 0,5s + 0,1 & 0,3s + 0,4 \end{bmatrix} \quad (43)$$

Figure 5 shows the plant's step response

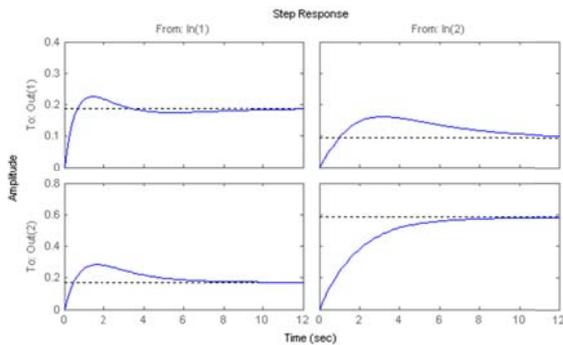


Figure 5: Step response of the controlled system

The matrix $M(s)$ on the right side of the diophantine equation (18) obtained from experiments is

$$M(s) = \begin{bmatrix} s^5 + 5s^4 + 10s^3 + & 0 \\ +10s^2 + 5s + 1 & \\ 0 & s^5 + 5s^4 + 10s^3 + \\ & +10s^2 + 5s + 1 \end{bmatrix} \quad (44)$$

The time responses of the control are shown in Figure 6.

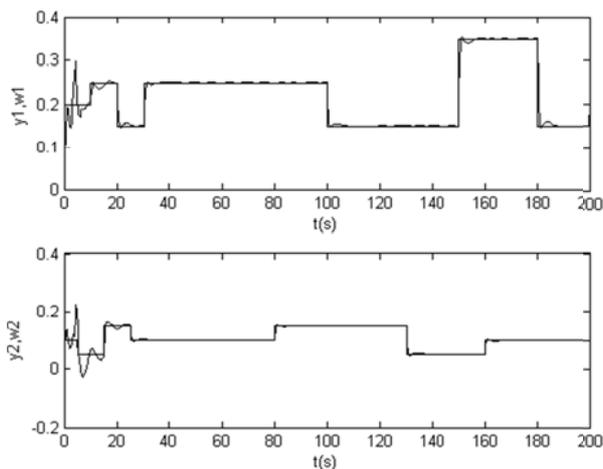


Figure 6: Adaptive control with decoupling controller

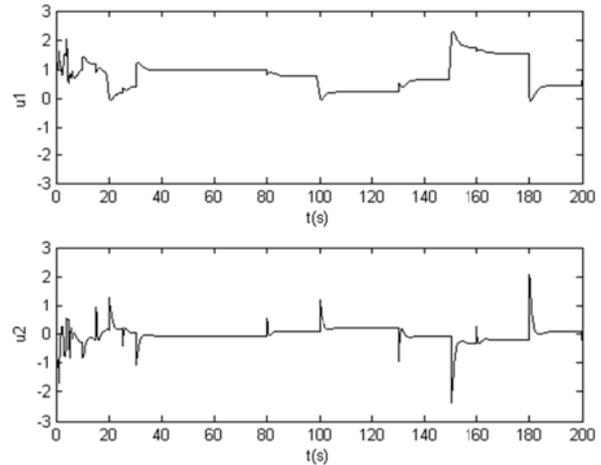


Figure 7: Adaptive control with decoupling controller-manipulated variables

From the courses of the variables in Figure 6 it is obvious that the basic requirements on control were satisfied. The system was stabilized and the asymptotic tracking of the reference signals was achieved. With regards to decoupling, interactions between the control loops are negligible.

CONCLUSIONS

A decoupling TITO controller was designed and implemented in the Matlab/Simulink environment. The simulation results proved that the method is suitable for control of linear systems. With regards to decoupling, it is clear that the compensator reduces interactions between the control loops. The internal structure of the implemented controller enables its easy modification and implementation of further similar control algorithms.

REFERENCES

- Bobál, V., J. Böhm, J. Fessler, J., Macháček. 2005. *Adaptive control*, Springer - Verlag, London.
- Cui H., E. W. Jacobsen. 2002. „Performance limitations in decentralized control“. *Journal of Process Control*, 12, 485–494.
- Chien I.L., D.E. Seborg and D. A. Mellichamp. 1987. “Self-Tuning Control with Decoupling”. *AIChE J.*, Vol. 33, No. 7, 1079 – 1088.
- Kubalčík, M. and V. Bobál. 2006. „Adaptive Control of Coupled Drives Apparatus Based on Polynomial Theory”. In Proc. IMechE Vol. 220 Part I: *J. Systems and Control Engineering*, 220(17), 641-654.
- Krishnawamy P.R. et al. 1991. “Reference System Decoupling for Multivariable Control”. *Ind. Eng. Chem. Res.*, 30, 662-670.
- Kučera, V. 1980. „Stochastic multivariable control: a polynomial approach“. *IEEE Trans. of Automatic Control*, 5, 913–919.
- Kučera, V. 1991. *Analysis and Design of Discrete Linear Control Systems*. Prentice Hall, Englewood Cliffs, New Jersey.

- Kučera, V. 1993. "Diophantine Equations in Control – a Survey". *Automatica*, 29, 1361 - 1375.
- Peng, Y. 1990. "A General Decoupling Precompensator for Linear Multivariable Systems with Application to Adaptive Control". *IEEE Trans. Aut. Control*, Vol. 35, No. 3, 344-348.
- Tade M. O., M.M. Bayoumi, D.W. Bacon. 1986. "Adaptive Decoupling of a Class of Multivariable Dynamic Systems Using Output Feedback". *IEE Proc. Pt.D*, No. 6, 265-275.
- Vidyasagar, M. 1985. *Control System Synthesis: A Factorization Approach*. MIT Press, Cambridge MA.
- Wahlberg, B. 1990. "On the Identification of Continuous – Time Dynamical Systems". *Report LiTH-ISY-I-0905*, Linköping.

AUTHOR BIOGRAPHIES



MAREK KUBALČÍK graduated in 1993 from the Brno University of Technology in Automation and Process Control. He received his Ph.D. degree in Technical Cybernetics at Brno University of Technology in 2000. From 1993 to 2007 he worked as senior lecturer at the Faculty of Technology, Brno University of Technology. From 2007 he has been working as an associate professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. Current work cover following areas: control of multivariable systems, self-tuning controllers, predictive control. His e-mail address is: kubalcik@fai.utb.cz.



VLADIMÍR BOBÁL graduated in 1966 from the Brno University of Technology, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are adaptive control and predictive control, system identification and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.

USE OF DYNAMIC MATRIX CONTROL IN SIMULATION OF HEAT SYSTEM

Stanislav Talaš², Vladimír Bobál^{1,2} and Adam Krhovják²

Tomas Bata University in Zlin

¹Centre of Polymer Systems, University Institute

²Department of Process Control, Faculty of Applied Informatics

T. G. Masaryka 5555

760 01 Zlin

Czech Republic

E-mail: talas@fai.utb.cz

KEYWORDS

Model Predictive Control, Dynamic Matrix Control, MATLAB/SIMULINK, Simulation, Heat Exchanger.

ABSTRACT

This paper demonstrates the use of Model Predictive Control (MPC) to system control. Dynamic Matrix Control (DMC) method was chosen and its functionality was verified by a simulation of system control based on a real laboratory model. A control algorithm and the simulation were realized in MATLAB/SIMULINK program environment. Results have proven capabilities of DMC method to control stable oscillatory and non-minimum phase systems. Additionally, real model parameters were tested with a demonstration of a possibility of tuning by a ratio of weighting values from objective function.

INTRODUCTION

Considering the scientific area of process control, it targets at present tendency of satisfying demands of the maximal productivity of the highest quality products at the lowest cost possible. With the power of the modern computing technology an approach of finding optimal results in reasonable time was made possible.

Advanced methods popular in industries with slow and large dimensional systems are predictive control methods (Qin & Badgwell, 2003). These techniques commonly contain an internal model for system behavior predictions. Gained information is further used to calculate a sequence of control inputs by minimizing a sum of squares between the desired and predicted trajectories. Therefore an optimal output is received in reference to the minimal error, eventually to the change of control inputs.

Development in this area started in 1980s with the publication of DMC method (Cutler & Ramaker, 1980). Original purpose of DMC was focused on multivariable constrained control problems, mainly occurring in chemical and oil industry. The influence of DMC caused its widespread use in world's major industrial companies (Morari & Lee, 1999).

Over the time there was a vast development of the DMC algorithm, its modifications and possibilities of application. (Garcia & Morshedi, 1986) provided a utilization

of a quadratic algorithm for an efficient handling of constraints, tuning and robustness. (Shridhar & Cooper, 1997) suggested a tuning strategy of DMC parameters for SISO systems, followed by an approach in case of MIMO systems (Shridhar & Cooper, 1998). (Dougherty & Cooper, 2003) described an approach to tune the parameters of the basic DMC algorithm for the case of integrating processes. In occurrence of nonlinear processes (Dougherty & Cooper, 2003) suggested a new adaptive control strategy using the output of multiple linear DMC controllers to maintain the performance over a wide range of operational levels.

The purpose of this paper is to give an insight on abilities of DMC options for the control of stable processes, primarily in the area of tuning its performance by changing the weight ratio of the optimization process between the output error and a demand of action value.

The paper is organized in the following way. General principles of MPC are presented first, followed by the description of specific properties of the DMC method. Basic functionality and characteristics are introduced. The final section presents the implementation of the simulation into the MATLAB/SIMULINK program environment and its results.

MODEL PREDICTIVE CONTROL

Predictive control is an approach to control a process through optimization. The main principle is in the prediction of future process outputs based on the inner model of the process. The goal of the control algorithm is to find such a vector of input values that the output of the model is optimal along the defined time area called horizon. To ensure robustness and stability an approach using feedback called the receding horizon strategy is often applied. From the vector of input values only the first value is used as an increment $\Delta u(k)$ added to the previous input giving the current input value $u(k)$. In the next step the entire procedure is repeated with new process output values.

The area of the optimization is defined by values of horizons representing the amount of sampling periods from the current time into the future. Values of horizons N_1 and N_2 limit the area, where the divergence between the desired and the output value is minimized. The horizon N_u limits the distance of steps where the action value is minimized.

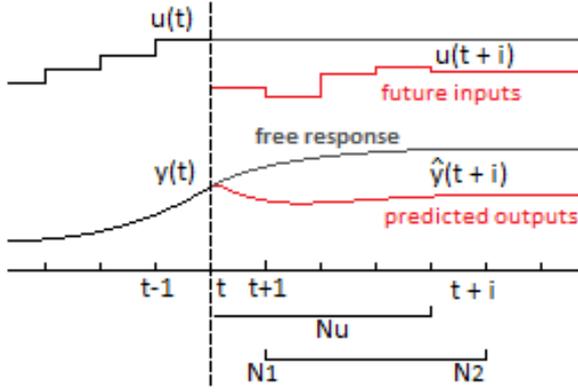


Figure 1: Receding horizon strategy

Calculation of the optimal output consists of a free response prediction describing the system behavior in the case of a constant input and the forced response with a reaction on a suggested series of inputs. Based on the superposition principle, the sum of these responses results in the future output prediction.

Several methods of MPC are used in practice; the main differences are in the description of the controlled process and in the objective function.

The Figure 2 shows a layout of the predictive control and a data transfer.

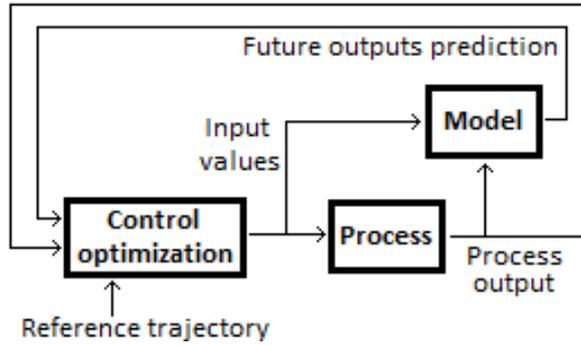


Figure 2: Basic structure of model predictive control

The optimization process is based on the minimization of values involved in control. Their mutual relations are formed by an objective function. The general expression of an objective function is

$$J = \sum_{i=N_1}^{N_2} \delta(i) [\hat{y}(k+i) - w(k+i)]^2 + \sum_{i=1}^{N_u} \lambda(i) [\Delta u(k+i-1)]^2, \quad (1)$$

where $\delta(i)$ and $\lambda(i)$ are weighting values, usually constants representing a ratio of the minimization between a divergence of output from the desired value and a change of the action value (Normey-Rico & Camacho, 2007).

Dynamic matrix control

DMC method is based on a step function limited to first N points. It is assumed that the process is stable and the disturbance prediction is stable all over the horizon

equal to actual difference between measured value $y_m(k)$ and its estimation $\hat{y}(k)$.

This method is often used in industry; some of the main reasons are an easy identification of the internal model and a possible addition of constraints into the minimization criterion. The disadvantage is an unsuitable use for unstable systems due to the step function in the internal model.

Predicted values along the horizon are given by the following equation

$$\hat{y}(k+i) = \sum_{j=1}^{N_u} g_j \Delta u(k+i-j) + \hat{n}(k+i), \quad (2)$$

where g_j represents the system step response value in time j and disturbance predictions $\hat{n}(k+i)$ are considered constant along the horizon.

By defining the free response

$$f(k+i) = y(k) + \sum_{j=1}^{N_u} (g_{i+j} - g_j) \Delta u(k-j), \quad (3)$$

predicted values are computed as

$$\hat{y}(k+i) = \sum_{j=1}^i g_j \Delta u(k+i-j) + f(k+i). \quad (4)$$

By expanding expression (4) to the number of future values given by the control horizon N_u an equation for the prediction of future outputs is created

$$\hat{\mathbf{y}} = \mathbf{G}\mathbf{u} + \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1, \quad (5)$$

where

$$\begin{aligned} \hat{\mathbf{y}}^T &= [\hat{y}(k+1) \dots \hat{y}(k+N_u)], \\ \mathbf{u}^T &= [\Delta u(k) \Delta u(k+1) \dots \Delta u(k+N_u-1)], \\ \mathbf{u}_1^T &= [\Delta u(k-1) \Delta u(k-2) \dots \Delta u(k-N_u)], \\ \mathbf{y}_1 &= \hat{y}(k). \end{aligned} \quad (6)$$

In equation (5) \mathbf{S} is a unitary vector with size $N_u \times 1$ and \mathbf{G} , \mathbf{H} are matrices with size $N_u \times N_u$

$$\mathbf{G} = \begin{bmatrix} g_1 & 0 & \dots & 0 \\ g_2 & g_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_{N_u} & g_{N_u-1} & \dots & g_1 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} (g_2 - g_1) & (g_3 - g_2) & \dots \\ (g_3 - g_1) & (g_4 - g_2) & \dots \\ \vdots & \vdots & \ddots \\ (g_{N_u+1} - g_1) & (g_{N_u+2} - g_2) & \dots \\ \dots & (g_{N_u+1} - g_{N_u}) \\ \dots & (g_{N_u+2} - g_{N_u}) \\ \vdots & \vdots \\ \dots & (g_{2N_u} - g_{N_u}) \end{bmatrix}. \quad (7)$$

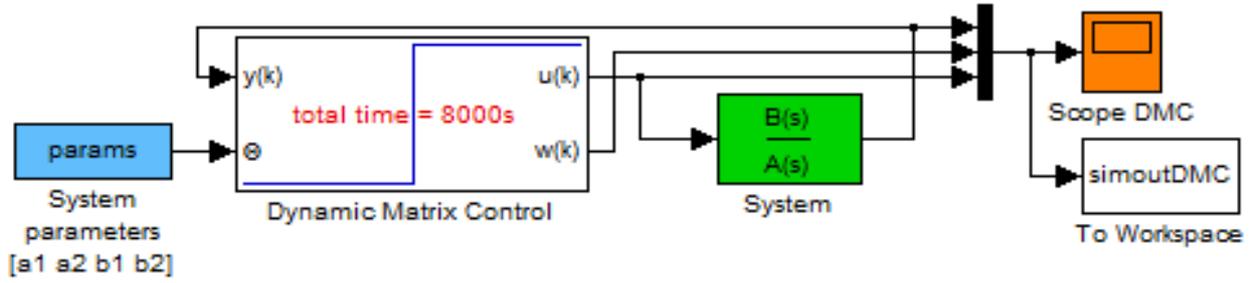


Figure 3: Simulation scheme in SIMULINK program environment

The output prediction is calculated using the sum of the free response and the forced response like

$$\hat{y} = \mathbf{G}\mathbf{u} + \mathbf{f}, \quad (8)$$

where \mathbf{f} is the free response given by equation

$$\mathbf{f} = \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1. \quad (9)$$

Optimization is performed by finding the minimal value of the objective function (1) accomplished by a differentiation with respect to the vector of the action variable \mathbf{u} and equating to zero. The shape of the control law is

$$\mathbf{u} = (\mathbf{G}^T \mathbf{Q}_\delta \mathbf{G} + \mathbf{Q}_\lambda)^{-1} \mathbf{G}^T \mathbf{Q}_\delta (\mathbf{f} - \mathbf{w}). \quad (10)$$

Where matrices \mathbf{Q}_δ and \mathbf{Q}_λ represent the weighting values from the objective function and are formed by identity matrices with value of δ respectively λ on the main diagonal. From the calculated array only the first value $\Delta u(k)$ is used in the current step. In each of the following steps the calculation is repeated (Camacho & Bordons, 2004), (Haber et al., 2011).

EXPERIMENTAL LABORATORY HEAT EQUIPMENT

A scheme of the laboratory heat equipment (Pekař et al., 2009) is described in Figure 4. The heat transferring fluid (e. g. water) is transported using a continuously controllable DC pump (6) into a flow heater (1) with max. power of 750 W. The temperature of a fluid at the heater output T_1 is measured by a platinum thermometer. Warmed liquid then goes through a 15 meters long insulated coiled pipeline (2) which causes the significant delay (20 – 200 s) in the system. The air-water heat exchanger (3) with two cooling fans (4, 5) represents a heat-consuming appliance. The speed of the first fan can be continuously adjusted, whereas the second one is of on/off type. Input and output temperatures of the cooler are measured again by platinum thermometers as T_2 , respective T_3 . The platinum thermometer T_4 is dedicated for measurement of the outdoor-air temperature. The laboratory heat equipment is connected to a standard PC via technological multifunction I/O card MF 624. This card is designed for the need of connecting PC compatible computers to real world signals. The card is designed for standard data acquisition, control applications and optimized for use with Real Time Toolbox for SIMULINK. The MATLAB/SIMULINK environment was used for all monitoring and control functions.

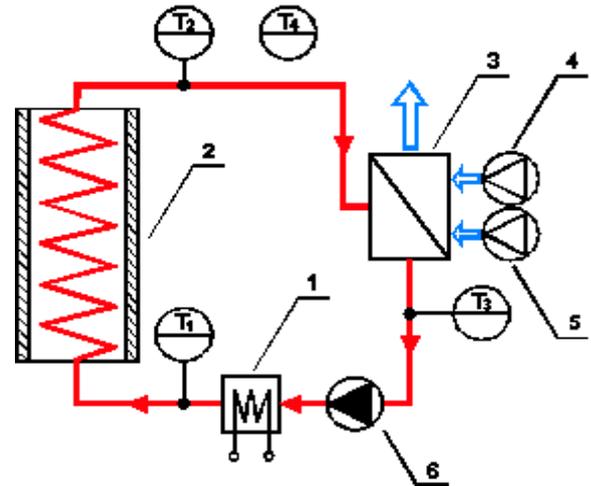


Figure 4: Scheme of laboratory heat equipment

SIMULATION CONTROL OF BASIC DYNAMICS

In order to verify the control algorithm a simulation scheme was created in the SIMULINK environment as can be seen in Figure 3. To test the general control capabilities of DMC, three systems representing different dynamics were chosen. These systems are described with following continuous transfer functions:

Stable non-oscillatory system

$$G_1(s) = \frac{2}{4s^2 + 5s + 1}.$$

Stable oscillatory system

$$G_2(s) = \frac{2}{4s^2 + 2s + 1}.$$

Non-minimum phase system

$$G_3(s) = \frac{-10s + 2}{4s^2 + 5s + 1}.$$

Discrete versions with sampling time $T_0 = 2s$ are

$$G_1(z^{-1}) = \frac{-0,7419z^{-1} + 0,0821z^{-2}}{1 + 0,4728z^{-1} + 0,2076z^{-2}}$$

$$G_2(z^{-1}) = \frac{-0,7859z^{-1} + 0,3679z^{-2}}{1 + 0,6806z^{-1} + 0,4834z^{-2}}$$

$$G_3(z^{-1}) = \frac{-0,7419z^{-1} + 0,0821z^{-2}}{1 - 1,0980z^{-1} + 1,7780z^{-2}}$$

Considering systems dynamics, the sizes of the control horizon N_u and the maximum horizon N_2 were set to 30 steps, while the minimum horizon N_1 remained 1. Control performances for corresponding systems are demonstrated with the desired trajectory with the shape of the step change known ahead by the control algorithm. Results can be seen in the following figures.

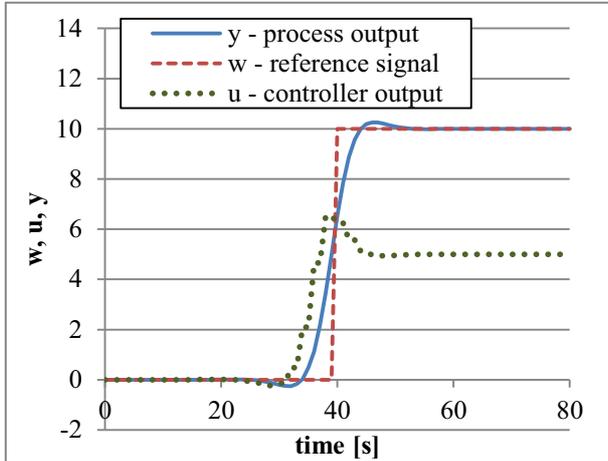


Figure 5: Control of stable non-oscillatory system G_1

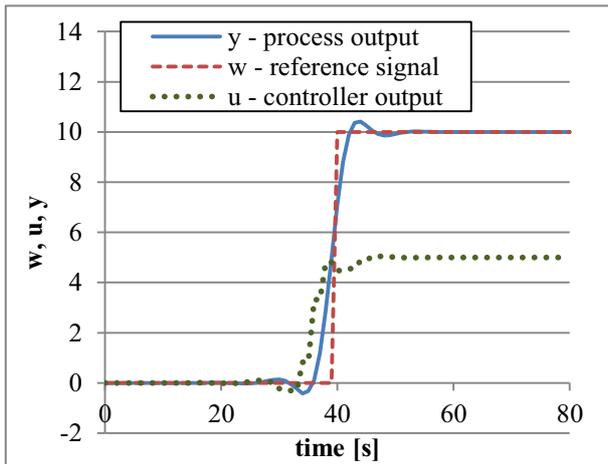


Figure 6: Control of stable oscillatory system G_2

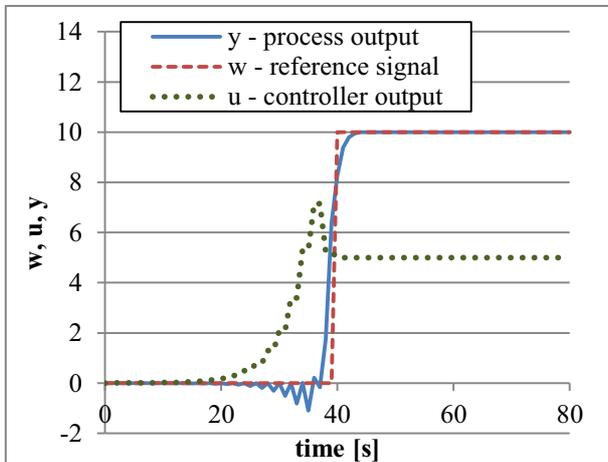


Figure 7: Control of non-minimum phase system G_3

Results have proven that the DMC algorithm is able to control stable, oscillatory and even non-minimum phase systems.

Simulation control of a heat system

The following simulation model is based on identification results of a laboratory heat-system. The real system is constructed as a circulation of water warmed by a heater and cooled by fans creating a stable circuit with large time constants. In order to concentrate the focus to the topic, the traffic delay caused by a fluid transport is not included in the following simulation.

This laboratory model was identified as a stable second order system with continuous expression

$$G_4(s) = \frac{0.001614s + 2.664 \cdot 10^{-5}}{s^2 + 0.03017s + 4.075 \cdot 10^{-5}}$$

and its discrete version with sampling time of 60 seconds

$$G_4(z) = \frac{0.0719z^{-1} - 0.0281z^{-2}}{1 - 1.097z^{-1} + 0.1636z^{-2}}$$

Values of horizons were set to 30 steps for the control horizon N_u as well as the maximum horizon N_2 , the minimum horizon N_1 was set to 1. As a tuning mechanism to gain a suitable precision with an appropriate action value, several different settings of weighting values $\delta(i)$ and $\lambda(i)$ ratio were examined.

The precision of the control process was determined by an integral of a squared error (ISE) criterion.

$$ISE = \int_0^{\infty} [e(t)]^2 dt \quad (11)$$

Due to large time constants of the model a step function was selected as a shape of the desired trajectory, in order to prove the ability of DMC to control the system optimally.

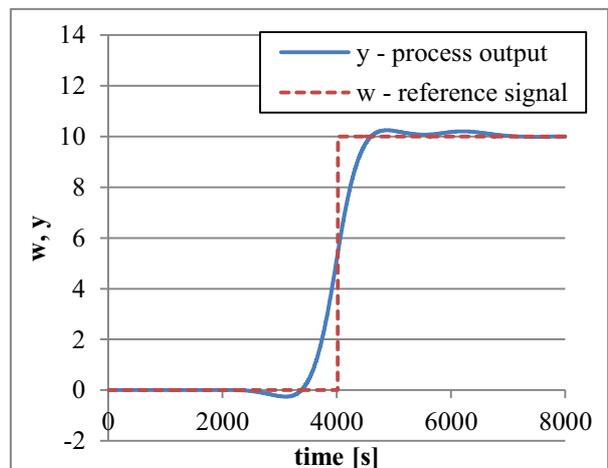


Figure 8: Reference and output values with weight parameters ratio 1:1

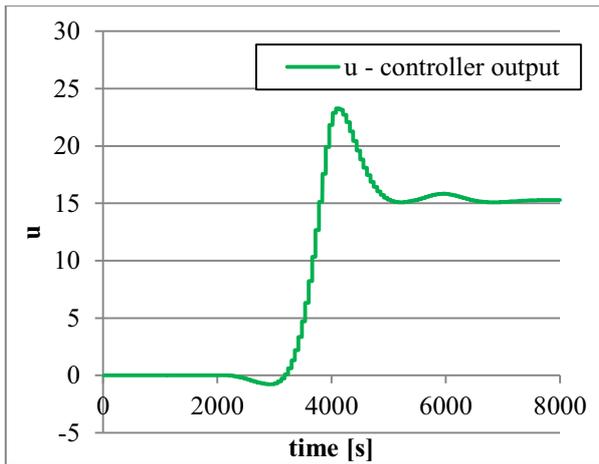


Figure 9: Controller output value with weight parameters ratio 1:1

Figures 8 and 9 demonstrate the control performance in the case of process optimization evenly distributed between the divergence of the output from the desired value and the change of the action value. This setting has achieved a value of the error criterion $ISE = 7005$.

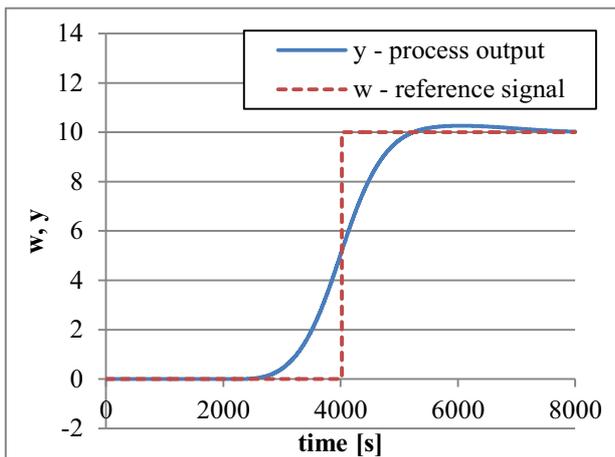


Figure 10: Reference and output values with weight parameters ratio 10:1

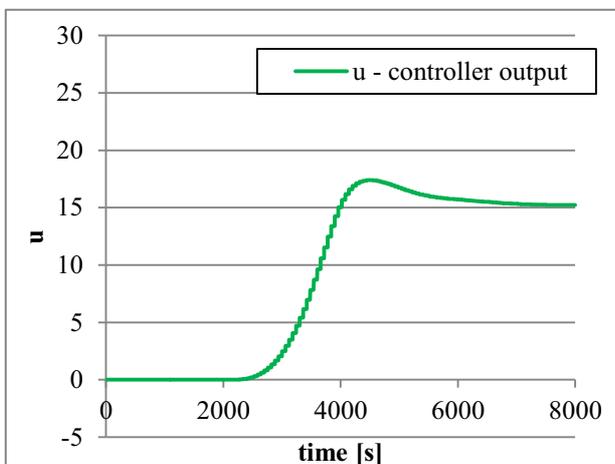


Figure 11: Controller output value with weight parameters ratio 10:1

Figures 10 and 11 illustrate the case of the increased weight parameter of the action value difference. Figure 11 clearly presents an increased smoothness of the controller output trajectory implying lower demands on a physical actuator. On the other hand, the precision of tracking the desired output value has significantly decreased, as can be seen in Figure 10, rising the value of the error criterion $ISE = 13390$.

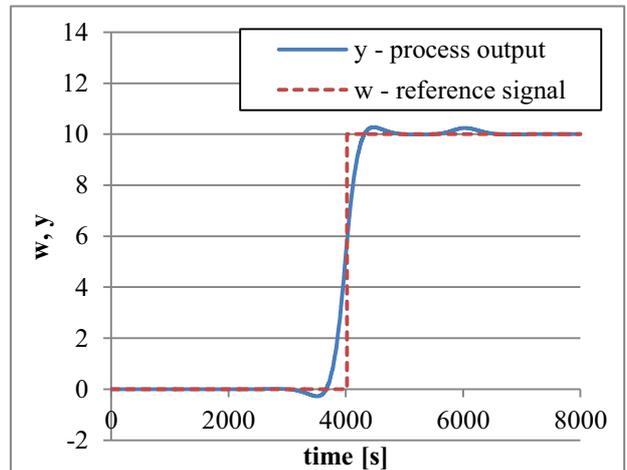


Figure 12: Reference and output values with ratio 1:10

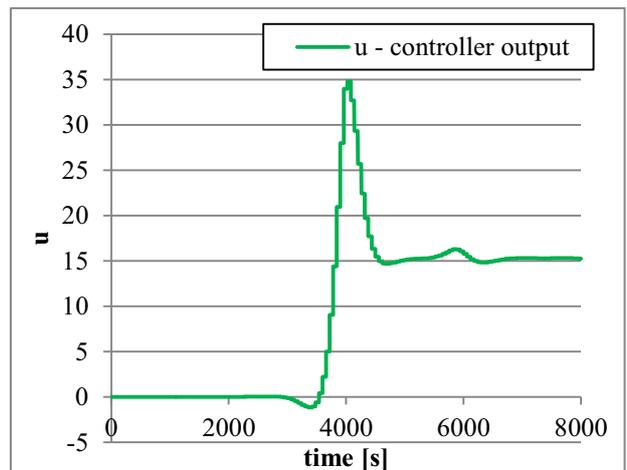


Figure 13: Controller output value with weight parameters ratio 1:10

Figures 12 and 13 show the case when the ratio is increased towards the divergence of the output from the desired trajectory, providing a high precision performance with a value of the error criterion $ISE = 3920$. This comes with a disadvantage of sharp changes in the action value and a considerably high peak value.

CONCLUSION

The paper presents results of the simulated control of the laboratory heat system by the DMC. Outcomes of simulations present the possibilities of tuning the performance of predictive control methods. A significant impact of the weight ratio was proven, as well as con-

nection between a precision of the output value and a dynamics of the action value as can be seen on Table 1.

Table 1: ISE comparison of weight parameters ratios

$\lambda : \delta$ weight ratio	ISE criterion value
1 : 1	7 005
10 : 1	13 390
1 : 10	3 920

Results have confirmed the ability of DMC to provide a high quality control, furthermore with use of weight values also to determine an amount of the required action value and the precision of the output value.

ACKNOWLEDGEMENT

This article was created with support of Operational Programme Research and Development for Innovations co-funded by European Regional Development Fund (ERDF), national budget of Czech Republic within the framework of the Centre of Polymer Systems project (reg. number: CZ.1.05/2.1.00/03.0111).

REFERENCES

- Camacho, E. F. & C. Bordons. (2004). Model Predictive Control, Springer-Verlag, London.
- Cutler, C. R. & B. L. Ramaker. (1980). Dynamic Matrix Control. *Proc. Joint Automatic Control Conference*, San Francisco, CA, paper WP5-B.
- Dougherty, D. & D. Cooper. (2003). A practical multiple model adaptive strategy for multivariable model predictive control. *Contr. Eng. Practice*, **11**, 649-664.
- Dougherty, D. & D. J. Cooper. (2003). Tuning guidelines of a dynamic matrix controller for integrating (non-self-regulating) proces. *Ind. Eng. Chem. Res.*, **42**, 1739-1752.
- Garcia, C. E. & A. M. Morshedi. (1986). Quadratic programming solution of dynamic matrix control (QDMC). *Chem. Eng. Commun.*, **46**, 73-87.
- Haber, R. R. Bars & U. Schmitz. (2011). Predictive Control in Process Engineering: From basics to the applications. Willey-VCH Verlag, Weinheim.
- Morari, M. & J. H. Lee. (1999). Model predictive control: past, present and future. *Computers & Chemical Engineering*, **23**, 667-682.
- Normey-Rico, J. E. & E. F. Camacho. (2007). Control of Dead-time Processes. Springer-Verlag, London.
- Pekař, L., R. Prokop & P. Dostálek. (2009). Circuit heating plant model with internal delays. *WSEAS Transactions on Systems and Control*, **8**, 1093-1104.
- Qin, S. J. & T. A. Badgwell. (2003). A survey of industrial model predictive control technology. *Control Engineering Practice*, **11**, 733-764.

Shridhar, R. & D. J. Cooper, (1997). A tuning strategy for SISO unconstrained model predictive control. *Ind. Eng. Chem. Res.*, **36**, 729.

Shridhar, R. & D. J. Cooper, (1998). A tuning strategy for unconstrained multivariable model predictive control. *Ind. Eng. Chem. Res.*, **37**, 4003-4016.

AUTHOR BIOGRAPHY



STANISLAV TALAŠ studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2013. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His e-mail address is talas@fai.utb.cz.



VLADIMÍR BOBÁL graduated in 1966 from the Brno University of Technology, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Slovak Republic. He is now Professor at the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. His research interests are adaptive and predictive control, system identification and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.



ADAM KRHOVJÁK studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2013. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interest focus on modeling and simulation of continuous time technological processes, adaptive and nonlinear control. He is currently working on programming simulation library of technological systems.

MULTIVARIABLE ADAPTIVE CONTROL OF TWO FUNNEL LIQUID TANKS IN SERIES

Adam Krhovják², Petr Dostál^{1,2} and Stanislav Talaš²

¹ Centre of Polymer Systems, University Institute, Tomas Bata University in Zlín,
Nad Ovčírnou 3685, 760 01 Zlín, Czech Republic.

² Department of Process Control, Faculty of Applied Informatics, Tomas Bata University in Zlín,
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
{krhovjak;dostalp;talas}@fai.utb.cz

KEYWORDS

Two funnel liquid tanks in series, nonlinear model, continuous-time model, input-output linear model, delta model, adaptive control, polynomial method, recursive identification, simulation software.

ABSTRACT

In this paper, we demonstrate continuous-time adaptive control for nonlinear system of two funnel tanks in series. For this purpose continuous-time input-output linear model is considered describing nonlinear system in the neighborhood of an operating point. The two approaches discussed here are direct estimation and alternate delta model. In the case of direct approach the key lies in the differential filters used to compute estimates. In the second strategy parameters of corresponding delta model are recursively estimated. A feedback control configuration with two degrees of freedom is adopted. In order to guarantee stability as well as asymptotic tracking of the step reference and step load disturbance attenuation, polynomial method is taken into account for the resulting controllers. A simulation software tool implementing illustrative model of two funnel liquid tanks in series has been studied in detail.

INTRODUCTION

In practice, most physical processes belong to a class of nonlinear systems. Although considerable research has been devoted to the SISO (Single Input – Single Output) nonlinear systems rather less attention has been paid to the MIMO (Multi Input – Multi Output) systems. Thus, we restrict our attention to the MIMO case.

As we move from linear to nonlinear systems superposition principle known from linear systems does not hold any longer and we are faced with more complicated situation. However, since linear models are so much more traceable, the first step in analyzing a nonlinear system comes with a trick of linearization. This is very intuitive approach but may run into problems trying to control system by using classical controller with a fixed parameters. Thus, the linearization alone is not sufficient. Here it is an idea to develop adaptive techniques eliminating limitations by introducing continuous-time input-output linear model (CT IOLM) with recursively estimated parameters. Thus, the control problem for the nonlinear system has been reduced to a

problem of designing adaptive controller for the nonlinear system.

A very considerable amount of effort has been spent in the control community in applying identification techniques to adaptive control. As a result of these experiments two classes of strategies are commonly utilized for learning unknown parameters of CT IOLM.

Since we assume that the reader is familiar with the theory of discrete-time linear systems, it is easy to see that if the sampling period converges to zero, then the parameters of the discrete model do not converge to their continuous counterparts. The way to deal with this trouble has been discussed by (Middleton and Goodwin 1990; Mukhopadhyay et al. 1992; Garnier and Wang 2008), coming with the strategy of an alternative discrete model known as delta. Even though they are considered to be discrete, parameter convergence is guaranteed as sampling period tends to zero. More advanced description is to be found in (Stericker and Sinha 1993). With this in mind, delta model is seen to be powerful enough for the purpose of identification.

On the other hand, there is also a large literature on identification procedure in which one can obtain estimates of the IOLM using the technique of filtering, e.g. (Young 1981; Wahlberg 1990). In this way, input-output variables are needed to go through the differential filters. However, in contrast to delta model strategy, it is not difficult to notice that this approach meets additional calculations, connected to the outputs of differential filters.

Throughout the paper, we have stressed to build an adaptive technique with two different identification strategies for nonlinear system of two funnel liquid tanks. In both cases, estimates of the system are, in turn, used to compute the controller parameters. Such a challenge has been met by organizing the control configuration with two degrees freedom, where both controllers have feedback form. We refer the interest reader to (Dostál et al. 2001; Ortega and Kelly 1984). Finally, we demonstrate a method of control law synthesis based on polynomial method (Kučera 1993; Mikleš and Fikar 2004) which ensures stability as well as asymptotic tracking of the reference signal. In addition, a great deal of effort has been spent on programming a nonlinear model of two funnel liquid tanks in series, bringing a practical simulation software tool.

MODEL OF TWO FUNNEL LIQUID TANKS

Consider the simple example of two funnel liquid tanks in series shown in Figure 2, where q_j and q_{jf} (for $j = 1, 2$) are outlet and inlet streams, respectively. Let us take D as the maximum diameter and H as the total height, which are same for both tanks. We define h_1 and h_2 as the liquid levels from the bottom.

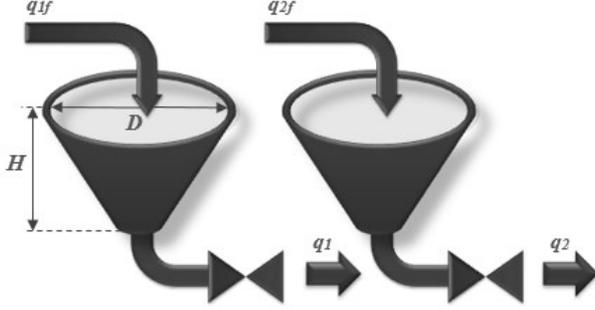


Figure 1: Two funnel liquid tanks in series

As an idealization, we can model them by the equations

$$\pi \frac{D^2}{4H^2} h_1^2 \frac{dh_1}{dt} + q_1 = q_{1f} \quad (1)$$

$$\pi \frac{D}{4H^2} h_2^2 \frac{dh_2}{dt} - q_1 + q_2 = q_{2f} \quad (2)$$

which are nonlinear. For a more detailed derivation we refer to (Corriou 1994; Bequette 2003). As the liquid moves through the valves, we see dependence of q_j on the liquid levels as

$$q_1 = k_1 \sqrt{|h_1 - h_2|} \quad (3)$$

$$q_2 = k_2 \sqrt{h_2} \quad (4)$$

where k_1, k_2 are valve constants.

The equilibrium points of the system are determined by setting $\dot{h}_1 = \dot{h}_2 = 0$ and solving for h_1 and h_2 . Therefore the equilibrium points correspond to the solution of

$$0 = \frac{4H^2}{\pi D^2 h_1^2} (q_{1f} - q_1) \quad (5)$$

$$0 = \frac{4H^2}{\pi D^2 h_2^2} (q_{2f} + q_1 - q_2) \quad (6)$$

Suppose $h_j \neq 0$ and $q_{jf} \neq 0$, and consider the change of variables

$$u_1(t) = q_{1f}(t) - \bar{q}_{1f}, \quad u_2(t) = q_{2f}(t) - \bar{q}_{2f} \quad (7)$$

$$y_1(t) = h_1(t) - \bar{h}_1, \quad y_2(t) = h_2(t) - \bar{h}_2 \quad (8)$$

In the new variables \mathbf{u} and \mathbf{y} , system has equilibrium at the origin. Therefore, in a small neighborhood of the origin, we may approximate the nonlinear system (1), (2) by its linearization about the origin. In particular, this implies that the continuous-time IOLM takes the form

$$\begin{bmatrix} s + a_{01} & a_{02} \\ a_{03} & s + a_{04} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} b_{01} & 0 \\ 0 & b_{04} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (9)$$

ADAPTIVE CONTROLLER DESIGN

In this section we will briefly discuss identification procedure and controller synthesis problem using polynomial approach.

Let us start by considering an adaptive control scheme depicted in Figure 1.

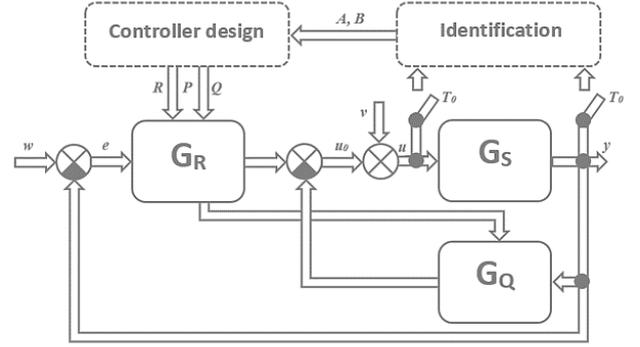


Figure 2: Control configuration

To give the reader a bird's eye view of the identification block, let us run two scenarios, answering the question of update law.

CT IOLM estimates

Since measurement of $\dot{\mathbf{y}}(t)$ and $\dot{\mathbf{u}}(t)$ is not available, we define vectors, representing filtered variables $\mathbf{y}_f(t)$ and $\mathbf{u}_f(t)$ by

$$\dot{y}_{1f} + c_0 y_{1f} = y_1, \quad \dot{y}_{2f} + c_0 y_{2f} = y_2 \quad (10)$$

$$\dot{u}_{1f} + c_0 u_{1f} = u_1, \quad \dot{u}_{2f} + c_0 u_{2f} = u_2 \quad (11)$$

From the form of ARX according to (Dostál et al. 2004) it is clear that the parameters may be estimated by

$$\dot{y}_{1f}(t_k) = b_{01} u_{1f}(t_k) - a_{01} y_{1f}(t_k) - a_{02} y_{1f}(t_k) + \varepsilon_1(t_k) \quad (12)$$

$$\dot{y}_{2f}(t_k) = b_{04} u_{2f}(t_k) - a_{03} y_{1f}(t_k) - a_{04} y_{2f}(t_k) + \varepsilon_2(t_k) \quad (13)$$

Delta model estimates

One of the difficulties in identifying ARX, proposed in previous subsection, is that we are required to have additional machinery associated with differential filters. As we will soon see, an alternative mechanism by which we identify parameters of the system, does not require any extra calculation.

For the sake of comparison let us consider competing delta model, which can be written as

$$\begin{bmatrix} \delta + \hat{a}_{01} & \hat{a}_{02} \\ \hat{a}_{03} & \delta + \hat{a}_{04} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \hat{b}_{01} & 0 \\ 0 & \hat{b}_{04} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (14)$$

Following the convention, the δ -operator is defined as

$$\delta = \frac{q-1}{T_0} \quad (15)$$

where q is the forward shift operator and T_0 corresponds to the sampling period.

This should come as no surprise in view of our earlier discussion on CT estimates that the equation (14) provides an essential factor used to formulate ARX model as follows

$$\delta y_1(\hat{k}-1) = b'_{01}u_1(\hat{k}-1) - a'_{01}y_1(\hat{k}-1) - a'_{03}y_2(\hat{k}-1) + \varepsilon_1(\hat{k}) \quad (16)$$

$$\delta y_2(\hat{k}-1) = b'_{04}u_1(\hat{k}-1) - a'_{02}y_1(\hat{k}-1) - a'_{04}y_2(\hat{k}-1) + \varepsilon_2(\hat{k}) \quad (17)$$

where the left-hand side of the above equations (16), (17) is given by

$$\delta y_i(\hat{k}-1) = \frac{y_i(\hat{k}) - y_i(\hat{k}-1)}{T_0} \quad (18)$$

Multivariable controller synthesis

Since the basis for the design of identification algorithms has been previously explored, we would like to find an adaptive controller that is a prescription for designing \mathbf{u} such that \mathbf{y} asymptotically tracks \mathbf{w} with all generated signals remaining bounded.

The model \mathbf{G} is a multi-input multi-output continuous-time IOLM, with transfer function matrix

$$\mathbf{G}(s) = \mathbf{A}^{-1}(s)\mathbf{B}(s) \quad (19)$$

To aid insight into controllers

$$\mathbf{G}_o(s) = \mathbf{Q}(s)\mathbf{P}^{-1}(s), \quad \mathbf{G}_r(s) = \mathbf{R}(s)\mathbf{P}^{-1}(s) \quad (20)$$

we will work with both reference signals and disturbance signals as follows

$$\mathbf{w}(s) = \begin{bmatrix} \frac{w_{10}}{s} & \frac{w_{20}}{s} \end{bmatrix}^T \quad (21)$$

$$\mathbf{v}(s) = \begin{bmatrix} \frac{v_{10}}{s} & \frac{v_{20}}{s} \end{bmatrix}^T \quad (22)$$

where $\mathbf{w}(s)$ stands for the references and $\mathbf{v}(s)$ is a vector of disturbances.

To proceed with the design of the controllers, we now restrict our attention to the question of deriving signals in control configuration. We leave it as an exercise for the reader to verify that (omit the dependence on s for simplifying notation)

$$\mathbf{y}(s) = \mathbf{A}^{-1}\mathbf{B}[\mathbf{u}_o(s) + \mathbf{v}(s)] \quad (23)$$

$$\mathbf{u}_o(s) = \mathbf{R}\mathbf{P}^{-1}[\mathbf{w}(s) - \mathbf{y}(s)] - \mathbf{Q}\mathbf{P}^{-1}\mathbf{y}(s) \quad (24)$$

$$\mathbf{y}(s) = \mathbf{P}\mathbf{D}^{-1}[\mathbf{B}\mathbf{R}\mathbf{P}^{-1}\mathbf{w}(s) + \mathbf{B}\mathbf{v}(s)] \quad (25)$$

$$\mathbf{e}(s) = \mathbf{P}\mathbf{D}^{-1}[(\mathbf{A}\mathbf{P} + \mathbf{B}\mathbf{Q})\mathbf{P}^{-1}\mathbf{w}(s) - \mathbf{B}\mathbf{v}(s)] \quad (26)$$

Looking at the substitution

$$\mathbf{D} = \mathbf{A}\mathbf{P} + \underbrace{\mathbf{B}(\mathbf{R} + \mathbf{Q})}_r \quad (27)$$

given in equations (25), (26) it can be seen that the control system is stable if we design \mathbf{D} to be Hurwitz.

From this point on, we concentrate our attention on showing how to ensure zero steady-state tracking error in the presence of uncertainties.

As is well known we have to use integral control such that

$$\mathbf{P}(s) = s\tilde{\mathbf{P}}(s) \quad (28)$$

$$\mathbf{Q}(s) = s\tilde{\mathbf{Q}}(s) \quad (29)$$

Substitution of these expressions back into the matrix Diophantine equation (27) yields

$$\mathbf{A}(s)s\tilde{\mathbf{P}}(s) + \mathbf{B}(s)\mathbf{T}(s) = \mathbf{D}(s) \quad (30)$$

where the polynomial matrix $\mathbf{T}(s)$ on the left hand side actually has the form

$$\mathbf{T}(s) = \mathbf{R}(s) + s\tilde{\mathbf{Q}}(s) \quad (31)$$

Another question that comes to our mind immediately is: What are the degrees of polynomial matrices \mathbf{T} , \mathbf{R} and $\tilde{\mathbf{Q}}$?

The rough answer to this question obviously takes the form

$$\deg \mathbf{T}(s) = \deg \mathbf{R}(s), \quad \deg \mathbf{T}(s) = \deg \mathbf{R}(s) \quad (32)$$

In order to give the reader sense of excitement let us uncover the partial solution of matrix Diophantine equation. Since the first term on the left hand side of (30) does not contain constant matrix of s^0 , we can observe that

$$\mathbf{B}_0\mathbf{T}_0 = \mathbf{D}_0, \quad \mathbf{R}_0 = \mathbf{T}_0 \quad (33)$$

Toward the goal, suppose we have succeeded in finding polynomial matrices

$$\mathbf{P}(s) = s\tilde{\mathbf{P}}(s) = \begin{bmatrix} sp_{01} & sp_{02} \\ sp_{03} & sp_{04} \end{bmatrix} \quad (34)$$

$$\mathbf{T}(s) = \begin{bmatrix} t_{11}s + t_{01} & t_{12}s + t_{02} \\ t_{13}s + t_{03} & t_{14}s + t_{04} \end{bmatrix} \quad (35)$$

that satisfy (30) for the diagonal matrix

$$\mathbf{D}(s) = \begin{bmatrix} (s + \alpha_1)^2 & 0 \\ 0 & (s + \alpha_2)^2 \end{bmatrix} \quad (36)$$

However, unfortunately this solution does not capture relationship that exist between matrices \mathbf{R} and \mathbf{Q} . As a guideline in searching for the relationship, the weighting matrix $\mathbf{\Gamma}$ is chosen. A large amount of practical experience tells us that in many cases it is reasonable to have weighting matrix in diagonal form.

$$\mathbf{\Gamma}(s) = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{12} \end{bmatrix} \quad (37)$$

Recalling the expansion of the polynomial matrix, see e.g. (Rosenwasser and Lampe 2006; Blomberg and Ylinen 2006), we can state the solution of the feedback control problem by polynomial matrices of the form

$$\mathbf{R}(s) = \begin{bmatrix} \gamma_{11}t_{11}s + t_{01} & \gamma_{11}t_{12}s \\ \gamma_{12}t_{13}s & \gamma_{12}t_{14}s + t_{04} \end{bmatrix} \quad (38)$$

$$\mathbf{Q}(s) = \begin{bmatrix} (1 - \gamma_{11}t_{11})s & (1 - \gamma_{11}t_{12})s \\ (1 - \gamma_{12}t_{13})s & (1 - \gamma_{12}t_{14})s \end{bmatrix} \quad (39)$$

So far, we have formed the basic idea of the control problem. All that remains now is to show that matrix feedback controllers take the form

$$\mathbf{G}_Q(s) = \begin{bmatrix} (1 - \gamma_{11})t_{11} & (1 - \gamma_{11})t_{12} \\ (1 - \gamma_{12})t_{13} & (1 - \gamma_{12})t_{14} \end{bmatrix} \quad (40)$$

$$\mathbf{G}_R(s) = \begin{bmatrix} \gamma_{11}t_{11} + \frac{t_{01}}{s} & \gamma_{11}t_{12} \\ \gamma_{12}t_{13} & \gamma_{12}t_{14} + \frac{t_{04}}{s} \end{bmatrix} \quad (41)$$

SIMULATION OF THE TWO FUNNEL LIQUID TANKS

We have developed a simulator that simulates adequately behavior of two funnel liquid tanks in series. Idealistic model has been implemented according to equations (1)-(4). The simulator has been coded in C# and can be used in different operation modes either to generate simulation data or run fast simulation of simplified model of two funnel liquid tanks. There are three essential components through which one can interact: a control initialization pane, an identification library pane, a chart pane.

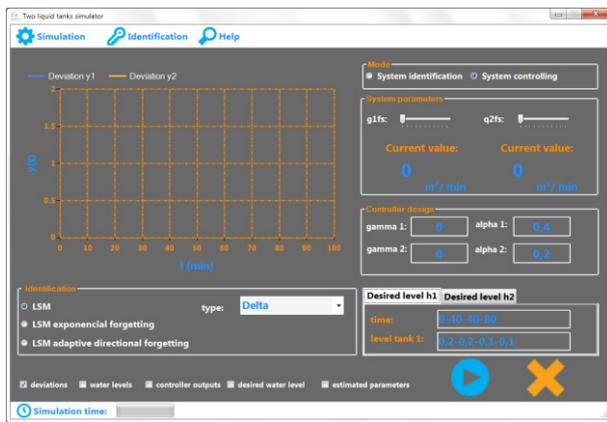


Figure 3: Two liquid tanks simulator

The simulator contains also several identification algorithms presented in the paper of (Kulhavý and Kárný 1984) which could be utilized for parameter estimation. The reference trajectory for each liquid level is entered in the simulator as a sequence of points representing combination of time values and liquid levels at corresponding sample times. The actual liquid levels of the tanks are then computed according two feedback controllers (40), (41).

SIMULATIONS AND RESULTS

In this section, we simulate the adaptive control of two funnel liquid tanks in series with the help of simulation tool, which was discussed in the previous section.

The parameters of the tanks and valves are $k_1 = 0.316 \text{ m}^{2.5}/\text{min}$, $k_2 = 0.296 \text{ m}^{2.5}/\text{min}$, $D = 1.5 \text{ m}$, $H = 2.5 \text{ m}$. The initial conditions we started with are $\bar{h}_1 = 1.8 \text{ m}$, $\bar{h}_2 = 1.4 \text{ m}$, $\bar{q}_{1f} = 0.2 \text{ m}^3/\text{min}$, $\bar{q}_{2f} = 0.15 \text{ m}^3/\text{min}$.

The results are illustrated in Figures 4-10 for different values of γ_{11} and γ_{12} as well as for different values α_1 and α_2 . Each figure shows the line plot of the control responses to step change in reference trajectories. In all cases, the recursive estimation of model parameters was performed with the constant sampling period $T_0 = 0.1 \text{ min}$. Figure 4 and 5 illustrate that there is an insignificant difference between two identification approaches.

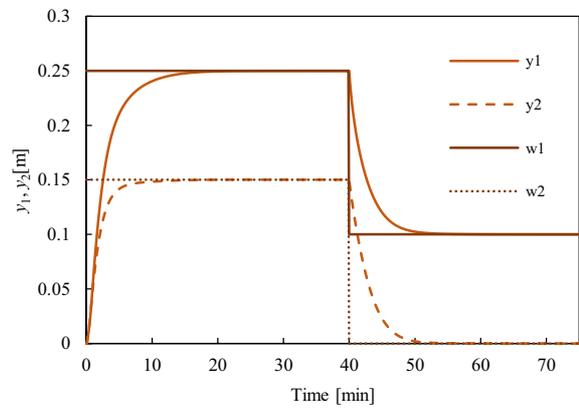


Figure 4: Controlled liquid levels – delta IOLM

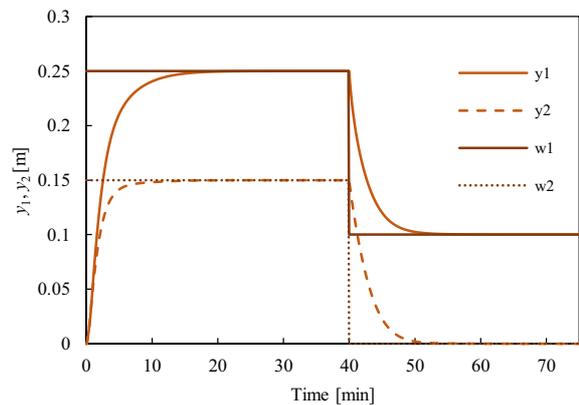


Figure 5: Controlled liquid levels – CT IOLM

This claim is also supported by Figure 6 in which parameter evolution of continuous-time and delta IOLM is captured. As can be seen from this example, we have found such a combination of parameters that gave a reasonably great responses with no overshoots.

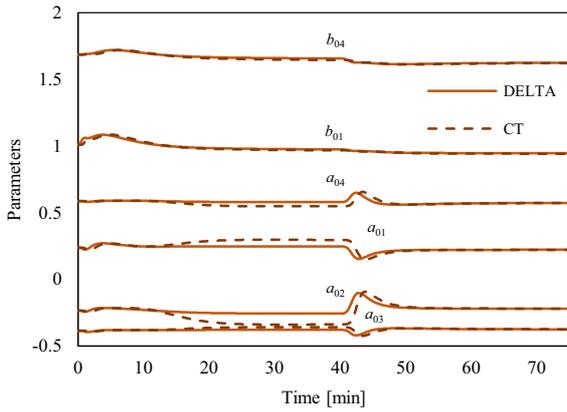


Figure 6: Parameter evolution during control

Figure 7 illustrates the significant and rapid effect of diagonal elements of weighting matrix Γ . Having higher values of γ_{11} and γ_{12} , we can accelerate the control. Unfortunately, their higher values greatly affect control input responses as is shown in Figure 8. This finding plays an important role in control design of real plants where the control inputs are physically limited.

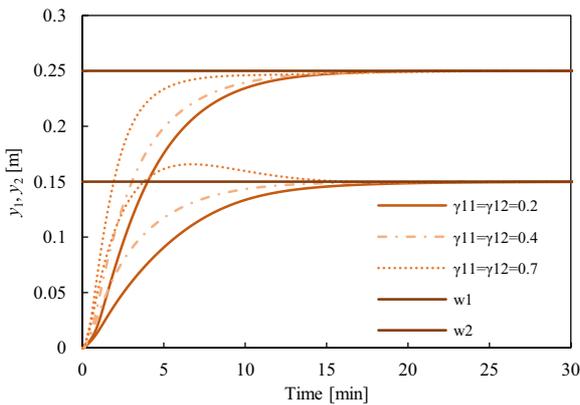


Figure 7: Controlled liquid levels – effect of γ_{11}, γ_{12}

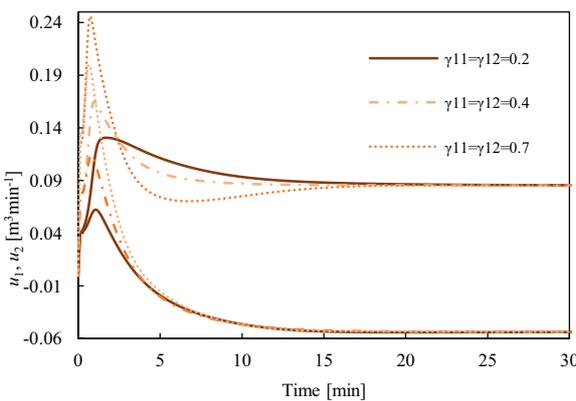


Figure 8: Control inlet flowrates – effect of γ_{11}, γ_{12}

Next simulation in Figure 9 shows the effect of double real pole α on the controlled output responses. It is easy to see that if we select lower values of poles in the open left-half plan, we speed up responses. However, one should be careful about experimenting with poles,

because too small poles usually correspond to overshoots. In other words, the system may go unstable before we have time to react. If we make a change in either the first pole or the second pole, then this will generally affect all the outputs, that is, there is interaction between the inputs and outputs. However, since we manipulate weights, there is minimal interaction between them. This important feature could be seen in Figure 10.

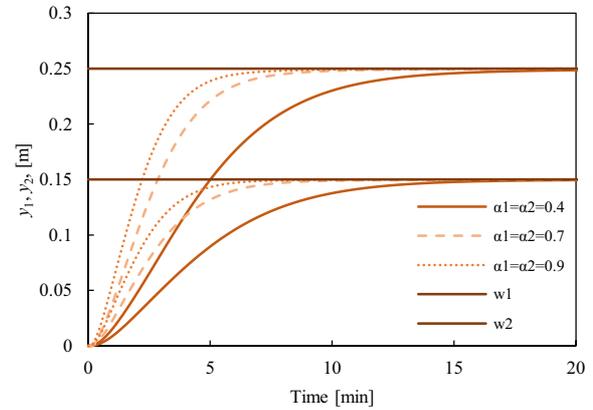


Figure 9: Controlled liquid levels – effect of α_1, α_2

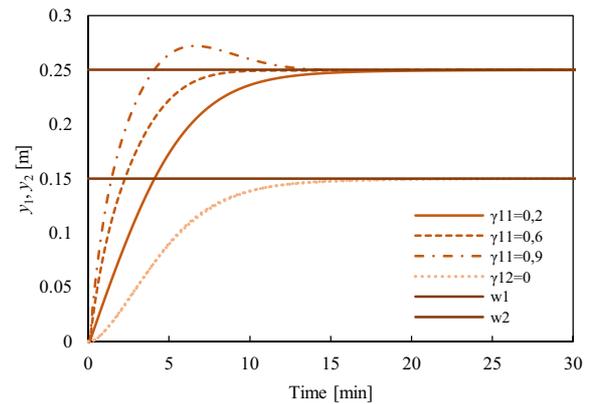


Figure 10: Controlled liquid levels – effect of γ_{11}

CONCLUSION

In this paper we illustrated a promising adaptive strategy to control nonlinear system of two funnel liquid tanks in series. The main motivation for our approach is to enable the use of design via linearization. We have demonstrated that if one uses linear model with some possibly time-varying parameters, control of nonlinear system may works suprisingly well. We have detailed identification problem of IOLM in two possible directions. In the first place, the direct estimation has been discussed in the sense of filtered variables. In the second place, we have applied alternative delta model approach to the estimation problem. The application has led to insight into the parameters of IOLM. As our results show, the algebraic procedure, based on the solution of the matrix Diophantine equation, is well suited for the controller synthesis. An adaptive control approach parameterizes the uncertainty in terms of certain unknown parameters of the IOLM matrices and tries to use feedback to learn

these parameters on line, that is, during the control of the nonlinear system. Tuning parameters of controller have a direct and major effect on feedback performance. However, finding the right combinations of tuning parameters to get prescribed behavior may be difficult. Much effort has been spent on the code development of reusable model of nonlinear system of two funnel liquid tanks in series and special care has been taken towards its implementation into simulation editor.

We are open to any suggestion and recommendation to improve our proposal.

ACKNOWLEDGEMENT

This article was created with support of Operational Program Research and Development for Innovations co-funded by European Regional Development Fund (ERDF), national budget of Czech Republic within the framework of the Centre of Polymer Systems project (reg. number: CZ.1.05/2.1.00/03.0111).

REFERENCES

- Bequette, B. 2003. *Process control: modeling, design, and simulation*. Prentice Hall. Upper Saddle River, N.J
- Blomberg, H. and R. Ylinen. 2006. *Algebraic theory for multivariable linear systems: modeling, design, and simulation*. Academic Press. New York.
- Corriou, J.-P. 2004. *Process control. Theory and applications*. Springer – Verlag, London.
- Dostál, P., V. Bobál and M. Blaha. 2001. “One approach to adaptive control of nonlinear processes”. In: *Proc. IFAC Workshop on Adaptation and Learning in Control and Signal Processing ALCOSP*, Cernobbio-Como, Italy, 407-412.
- Dostál, P., V. Bobál, and F. Gazdoš. 2004. “Adaptive control of nonlinear processes: Continuous-time versus delta model parameter estimation”. In: *Proceedings of 8th IFAC Workshop on Adaptation and Learning in Control and Signal Processing ALCOSP 04*, Yokohama, Japan, 273-278.
- Garnier, H. and L. Wang (eds.). 2008. *Identification of continuous-time models from sampled data*. Springer-Verlag, London, 2008.
- Kučera, V. 1993. “Diophantine equations in control – A survey”. *Automatica*, 29, 1361-1375.
- Kulhavý, R. and M. Kárný. 1984. “Tracking of slowly varying parameters by directional forgetting”. In: *Proc. 9th IFAC World Congress*, vol. X, Budapest, 78-83.
- Middleton, R.H. and G.C. Goodwin. 1990. *Digital Control and Estimation - A Unified Approach*. Prentice Hall. Englewood Cliffs.
- Mikleš, J. and M. Fikar. 2004. *Process modelling, identification and control 2*, STU Press, Bratislava, Slovakia, 2004.
- Mukhopadhyay, S., A.G. Patra and G.P. Rao. 1992. “New class of discrete-time models for continuous-time systems”. *International Journal of Control*, 55, 1161-1187.
- Ortega R and R. Kelly. 1984. “PID self-tuners: Some theoretical and practical aspects”. *IEEE Trans. Ind. Electron.*, Vol. IE-31, 332-338.
- Rosenwasser, E. and B. P. Lampe. 2006. *Multivariable computer-controlled systems: a transfer function approach*. Springer – Verlag, London.
- Stericker, D.L. and N.K. Sinha. 1993. “Identification of continuous-time systems from samples of input-output data using the δ -operator”. *Control-Theory and Advanced Technology*, 9, 113-125.
- Wahlberg, B. 1988. “On the identification of continuous time dynamical systems”. In *Proc. IFAC symposium on Identification and Parameter Estimation*, Beijing, China.
- Young, P. C. 1981. “Parameter estimation for continuous time models – A survey”. *Automatica*, 17, 23-29.

AUTHOR BIOGRAPHIES



ADAM KRHOVJÁK studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2013. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interest focus on modeling and simulation of continuous time technological processes, adaptive and nonlinear control.



PETR DOSTÁL studied at the Technical University of Pardubice, where he obtained his master degree in 1968 and PhD. degree in Technical Cybernetics in 1979. In the year 2000 he became professor in Process Control. He is now Professor in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interest are modeling and simulation of continuous-time chemical processes, polynomial methods, optimal and adaptive control.



STANISLAV TALAŠ studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2013. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His e-mail address is talas@fai.utb.cz.

MODELING OF ALCOHOL FERMENTATION IN BREWING – CARBONYL COMPOUNDS SYNTHESIS AND REDUCTION

Vessela Naydenova, Vasil Iliev, Maria Kaneva, Georgi Kostov
Department “Technology of wine and brewing”

University of Food Technologies
4002, 26 Maritza boulvd., Plovdiv, Bulgaria

E-mail: george_kostov2@abv.bg; vesi_nevelinova@abv.bg; m_kaneva@abv.bg; vasil_iliev_1988@abv.bg;

PetiaKoprinkova-Hristova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Acad.G. Bonchev Str., 25-A, 1113 Sofia, Bulgaria
E-mail: pkoprinkova@bas.bg

Silviya Popova
Institute of System Engineering and Robotics
Bulgarian Academy of Sciences
Acad. G. Bonchev str. bl.2, Sofia-1113, Bulgaria
E-mail: popova_silvia2000@yahoo.com

KEYWORDS

modeling, kinetics, vicinal diketones, aldehydes, metabolism

ABSTRACT

A mathematical model was developed for studying the carbonyl compounds synthesis and reduction in beer fermentation with alginate-chitosan microcapsules with liquid core. The model was based on the results for the influence of the fermentation temperature, original wort gravity and immobilized cells mass on the carbonyl compounds synthesis and reduction. The obtained model described with high accuracy the vicinal diketones synthesis and reduction and confirmed the experimental data. However, the model was not in agreement with the data for aldehydes synthesis and reduction. It did not take into account the second peak in aldehyde concentration during maturation. It can be assumed that the peak was related to maltotriose uptake by the used yeast strain. Nevertheless, the obtained model can be used for the description of carbonyl compounds synthesis and reduction in beer fermentation with immobilized cells.

INTRODUCTION

In brewing, the fermentation is one of the longest stages as well as an important aromatic compound production step. Indeed, fermentation has the main impact on process productivity and product quality. The brewing process productivity can be increased by the introduction of immobilized cells technology. It allows beer production to be accomplished in as little as 2-3 days (Branyik et al, 2005).

Yeast metabolism during fermentation and maturation affects significantly beer flavor. Ethanol, CO₂, esters and fusel alcohols have positive contributions to beer flavor. Dimethyl sulphide, hydrogen sulphide, and carbonyl compounds contribute to beer flavor defects (Meilgaard, 1975).

Carbonyl compounds are important because they have a high flavor potential and a significant influence on the

flavor stability of beer. Over 200 carbonyl compounds have been detected in beer (Rusell, 2006). The most important carbonyls are acetaldehyde and VDK. Acetaldehyde has unpleasant “grassy” flavor and aroma. It is of special interest because of its role as the immediate precursor of ethanol. VDK – diacetyl (2,3-butanedione) and 2,3-pentanedione have “butterscotch” and “toffee” aroma and taste (Briggs et al., 2004). The VDK concentrations in beer determined the maturation process time (Wilaert, 2007).

The aim of this work was to develop a mathematical model of the carbonyl compounds synthesis and reduction in beer fermentation with immobilized yeast. The model parameters identification was based on experimental data for the effect of T_{MF}, OE and Mic on the synthesis and reduction of VDK and aldehydes in beer produced under laboratory conditions.

MICROORGANISMS AND FERMENTATION CONDITIONS

The fermentations were carried out with bottom-fermenting yeast strain *Saccharomyces pastorianus* Saflager S-23. Immobilization procedure was previously reported in (Parcunev et al., 2012). The fermentations (main and secondary) were carried out with 400 cm³ sterile wort in fermentation bottles equipped with airlocks. The fermentation conditions are shown in Table 1. The data was part of planned experiment schedule which was reported in (Naydenova et al., 2014). The maturation temperature was 4°C higher than the T_{MF}. Maturation started when the difference between the attenuation limit and apparent attenuation was approximately 20% (Naydenova et al, 2014). The characterization of wort, green beer and beer (OE, degree of attenuation, extract, alcohol and VDK) was conducted according to the current methods recommended by the European Brewery Convention (Analytica-EBC, 2004). The aldehyde concentrations were determined according to (Marinov, 2010). Biomass concentration in immobilized cells was determined according to the mathematical model proposed in (Parcunev et al., 2012).

PARAMETERS IDENTIFICATION

The fermentation process kinetics was described with ordinary differential equations system (1).

$$\begin{aligned} \frac{dX}{dt} &= \mu(t, T) X(t, T) \\ \frac{dP}{dt} &= q(t, T) X(t, T) \\ \frac{dS}{dt} &= -\frac{1}{Y_{x/s}} \frac{dX}{dt} - \frac{1}{Y_{p/s}} \frac{dP}{dt} \\ \frac{dVDK}{dt} &= Y_{VDK} \cdot \mu(t, T) \cdot X(t, T) - k_{X, VDK} \cdot VDK(t, T) \cdot X(t, T) \\ \frac{dA}{dt} &= Y_A \mu(t, T) X(t, T) - k_A A(t, T) X(t, T) \\ \mu &= \mu_{\max} \frac{S}{K_{sx} + S}; \quad q = q_{p\max} \frac{S}{K_{sp} + S} \end{aligned} \quad (1)$$

The parameters identification was made by software programs in MatLab Environment (Kostov et al., 2012; Mitev and Popova, 1995; Popova 1997). The software minimized the sum of squared of difference between the model outputs and experimental data with respect of models parameters:

$$E(v) = (X(k_1, k_2, \dots, k_n) - X^e)^2 + (S(k_1, k_2, \dots, k_n) - S^e)^2 + (P(k_1, k_2, \dots, k_n) - P^e)^2 + (VDK(k_1, k_2, \dots, k_n) - VDK^e)^2 + (A(k_1, k_2, \dots, k_n) - A^e)^2 \quad (2)$$

For that purpose the function “fmincon” was applied. Here k_i , $i=1 \div n$ was model parameters vector which has to be determined as minimization procedure output. For that purpose the following complimentary differential equation:

$$dk_i / dt = 0 \quad (3)$$

was added to the ordinary differential equations model because k_i , $i=1 \div n$ was constant. For solving the overall differential equations system based on the Runge-Kutta formula of 4-5 order was used MATLAB function “ode45”. All parameters are shown in table 3.

CARBONYL COMPOUNDS SYNTHESIS AND REDUCTION

Vicinal diketones

Diacetyl and 2,3 - pentanedione are produced by yeast during fermentation. Diacetyl is the more important substance because of its lower flavor threshold. Both VDK are formed from intermediates of the amino acid biosynthesis. Diacetyl relates to valine and 2,3-pentanedione relates to isoleucine. The first intermediates in this metabolism are α -acetylacetyl and α -acetohydroxybutyrate. These components are discharged from the cell and undergo an oxidative decarboxylation to form diacetyl and 2,3-pentanedione. Yeast takes in these substances again and reduces them to 2,3-butanediol and 2,3-pentanediol, respectively. Owing to their high threshold, both resulting components show little influence on flavor (Handbook of brewing: Processes, Technology, Markets 2009; Debourg, 1999). The formation of the diketones is illustrated in Figure1.

Table 1
Fermentation conditions for beer production with immobilized cells

№	T _{MF}	T _{MATF}	OE	M _{IC}
-	°C	°C	% w/w	g
1	10	14	10.5	5
2	10	14	10.5	15
3	12.5	16.5	13	10
4	12.5	16.5	8.5	10
5	15	19	10.5	5
6	15	19	10.5	15
7	17	21	13	10

Aldehydes synthesis and reduction

Several aldehydes arise during wort production; others are formed as intermediates in the biosynthesis of higher alcohols from oxo-acids by yeasts (Briggs et al., 2004). Acetaldehyde synthesis is linked to yeast growth. Its concentration is maximal at the end of the growth phase, and is reduced at the end of the primary fermentation and during maturation by the yeast cells (Willaert, 2007). Removal of acetaldehyde is favored by increased yeast content during maturation (Russell, 2006)

MATHEMATICAL MODELS AND THEIR EXPLANATION

The fermentation with immobilized cells can be described with the equations for batch fermentation with free cells as previously reported (Parcunev et al., 2012; Vassilev et al., 2013). These are the first three equations in (1). For the adequate model development it is necessary to take into account some steps in the metabolites synthesis and reduction during beer fermentation.

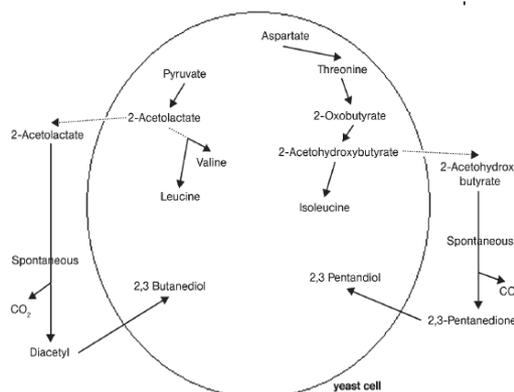


Figure 1. Formation and reduction of vicinal diketones
(Handbook of brewing: Processes, Technology, Markets 2009)

The VDK synthesis is a result of sugars and amino acids uptake by yeast for α -acetylacetyl and α -acetohydroxybutyrate production. Therefore, these two components have to be considered in the model. On the other hand, the sugars and amino acids uptake is associated with yeast growth. Thus, the VDK synthesis

is associated with cell growth. It can be taken into account in the model by a yield coefficient Y_{VDK} .

VDK reduction has two stages – chemical and biological. The first phase is the chemical conversion of α -acetolactate and α -acetohydroxybutyrate to diacetyl and 2,3-pentanedione, respectively. It can be intensified with the increase in temperature during maturation. The biological stage includes the uptake of diacetyl and 2,3-pentanedione by yeasts and their reduction to acetoin and 2,3-pentanediol, respectively. This phase can also be intensified with the increase in temperature. In our experiments the maturation temperature was 4 °C higher than the main fermentation temperature, which resulted in faster VDK reduction. VDK synthesis and reduction can be described with the fourth equation in the system (1) after considering all the factors that affect it.

In our previous study (Vassilev et al., 2013) it was shown that the aldehyde synthesis was associated with yeast growth, and their reduction - with the biomass concentration in the bioreactor and the aldehydes concentration in fermenting beer. These dependencies are presented in the fifth equation of the differential equations system (1).

According to differential equations system (1) the kinetic parameters depended on the fermentation temperature.

In the works of Andreas-Toro et. al, 1998 and Ramirez and Maciejowski, 2007 was found that the kinetic parameters and especially the specific growth rate could be described with an equation similar to the Arrhenius equation:

$$\mu_i = \mu_{i_{max}} \exp\left(-\frac{E_{\mu i}}{RT}\right) \quad (4)$$

Therefore, it can be suggested that the specific growth rate is a function of the cultivation conditions. Such kind of models are developed for the processes with free cells, but the diffusion resistances in the processes with immobilized cells may lead to some differences, which have to be considered. Thus, for simplification, the initial studies were made by the mathematical equations system and Monod equation (1). In the present work $E_{\mu i}$ was not evaluated. Ramirez and Maciejowski, 2007 showed that $E_{\mu i}$ ranged between -68.4 and 211.9 kcal/gmole. The value depended on the following parameters: the specific growth rate, specific substrate consumption rate or the specific metabolites production rate.

RESULTS OF FERMENTATIONS AND KINETIC PARAMETERS

Figure 2 and Figure 3 present the fermentation dynamics for one of the investigated variants (variant 3 of Table 1 and variant 7 of Table 1, respectively) as well as the comparison between the mathematical model (1) and the experimental data. The other variants showed similar fermentation dynamics.

The primary fermentation time and maturation time of the studied variants are presented in Table 2. It can be

found that the increase in the T_{MF} (the maturation temperature, respectively) led to the fermentation time reduction. It has to be noted that the observed trend deviation in the variant 7 was due to the fermentation kinetics.

The kinetic parameters of the studied fermentations are presented in Table 3. The results confirmed our previous observations (Parcunev et al., 2012; Vassilev et al., 2013) that the immobilization did not significantly affect the primary metabolism of immobilized yeast. The kinetic parameters indicated high specific fermentation rate (dX/dt and dP/dt), which decreased with the increase in OE due to substrate inhibition and catabolite repression. The maximum specific ethanol production rate varied between 0.48 and 1.19 g/(g.h) depending on the operational conditions. The major amount of the ethanol was produced by the end of the main fermentation because 80% of fermentable sugars were fermented during the primary fermentation.

Table 2
Fermentation time of experimental variants
(according Naydenova et al, 2014)

№	Time _{MF}	Time _{MATF}	Time
-	h		
1	288	168	456
2	192	204	456
3	204	180	294
4	108	156	264
5	144	96	240
6	78	156	234
7	120	172	292

The obtained results for the VDK synthesis and reduction were very interesting. The increase in T_{MF} resulted in an increase in the average VDK synthesis rate ($\mu \cdot X \cdot Y_{VDK}$). The most interesting results were recorded during the main fermentation at highest temperature (variant 7). The results obtained did not confirm the suggestion that the yield coefficient Y_{VDK} would be very high. It can be explained by the simultaneous VDK synthesis and reduction at 17 °C. M_{IC} increase affected contradictory the VDK synthesis. It should be noted that the M_{IC} increase led to an increase in yield coefficient Y_{VDK} at 10°C (variants 1 and 2). On the contrary, at 15 °C the M_{IC} increase resulted in decreased yield coefficient Y_{VDK} (variants 5 and 6). It can be assumed that the combination of high T_{MF} and M_{IC} caused accelerated VDK reduction, which took place simultaneously with VDK synthesis. At constant T_{MF} and M_{IC} the OE increase led to decrease in Y_{VDK} (variants 3 and 4).

The specific VDK reduction rate depends on the local VDK concentration and the biomass concentration. However, the increase in temperature and biomass concentration resulted in accelerated VDK reduction. Unfortunately, the VDK synthesis and reduction rates in the microcapsules could not be determined, because it was difficult to measure the VDK concentration in the capsule.

Figure 4 shows the average VDK reduction rate by the yeasts in stationary growth phase. It was calculated by the multiplication of VDK reduction coefficient ($K_{X,VDK}$) and the biomass concentration in stationary growth phase ($X(\text{stat})$). It can be found that there was a region with optimal operational conditions for carrying out maturation – OE=10-13 °P and maturation temperature 14-19 °C. Therefore, the optimal

fermentation conditions were OE=10-13 °P and T_{MF} = 10-15°C. This coincides with the optimal interval for fermentation using bottom-fermenting yeast stains. The increase in temperature resulted in deterioration in beer quality. The VDK reduction rate at temperatures above 21 °C was not investigated because these temperatures were not proper for lager beer maturation.

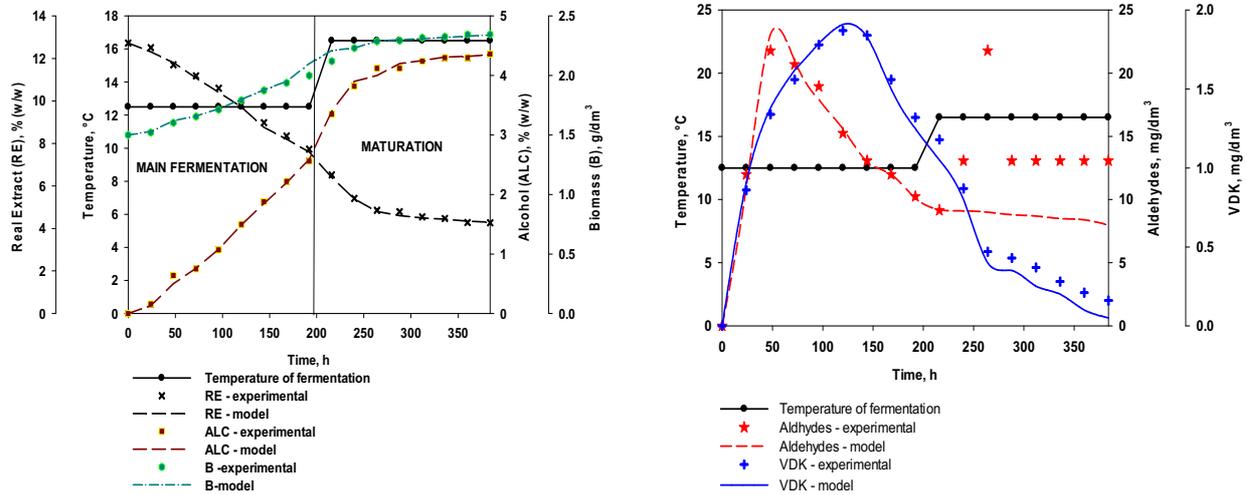


Fig.2. Fermentation dynamics(var. 3, tabl.1)

* experimental results according to Naydenova et al, 2014

Table 3

Kinetic parameters of the alcoholic fermentation and the carbonyl compounds synthesis and reduction

No	μ_{max} h ⁻¹	K_{SX} g/dm ³	q_{pmax} g/(g.h)	K_{SP} g/dm ³	$Y_{X/S}$ -	$Y_{P/S}$ -	Y_{VDK} mg/(g.h)	$K_{X,VDK}$ mg/(g.h)	Y_A mg/(g.h)	K_A mg/(g.h)	E(r)
1	0.158	229.5	1.19	229.5	0.148	13.83	2.80	0.012	39.1	0.0056	7.73
2	0.124	248.6	0.49	209.8	0.229	44.42	3.47	0.048	26.7	0.0011	5.57
3	0.421	240.5	0.53	228.2	0.051	28.42	5.34	0.091	39.8	0.0012	8.71
4	0.493	224.7	0.96	216.9	0.151	7.155	6.63	0.087	69.5	0.009	10.1
5	0.195	256.4	0.48	246.8	0.310	4.42	7.23	0.026	98.2	0.0226	4.3
6	0.278	241.2	0.51	246.7	0.13	5.81	4.48	0.025	100.3	0.125	5.2
7	0.387	245.6	1.09	249.6	0.015	6.21	4.54	0.022	105.2	0.109	10.2

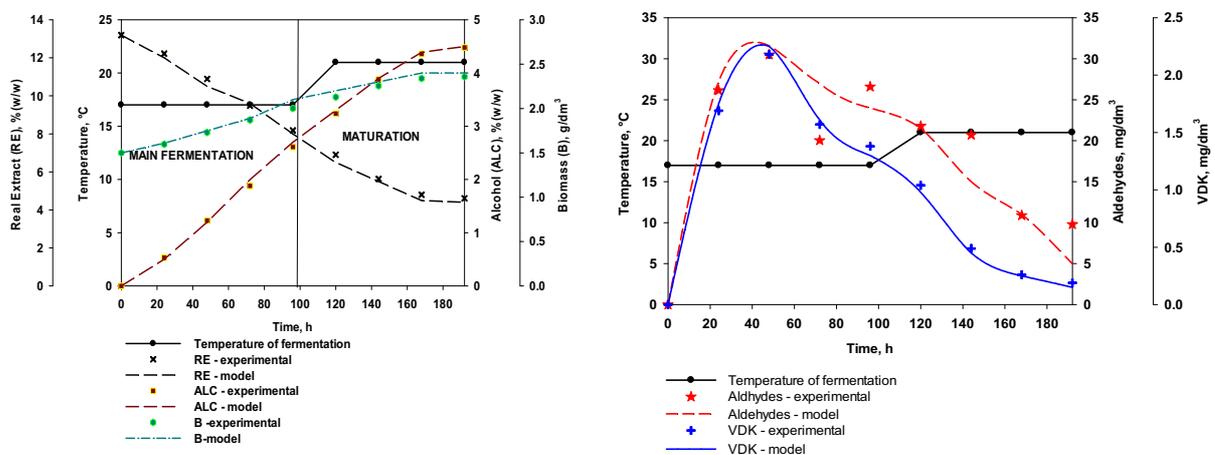


Fig.3. Fermentation dynamics(var. 7, tabl.1)

* experimental results according to Naydenova et al, 2014

It has to be noted that the different VDK synthesis rates caused the VDK maximums to be determined at different phases of main fermentation. At low T_{MF} and M_{IC} the maximum concentration was detected at the end of main fermentation. The increase in T_{MF} and M_{IC} resulted in VDK peaks at the beginning of the main fermentation (1-3 days). The data confirmed the observations in Naydenova et al., 2014.

The model showed almost complete diacetyl and 2,3-pentanedione reduction. Nevertheless, the VDK concentration in beer produced with immobilized yeast was higher than the VDK concentration in conventional beer.

Interesting trends were observed for the aldehyde synthesis and reduction (Table 3). The increase in T_{MF} , led to the accelerated aldehydes synthesis. It is interesting to note that the T_{MF} increase with 5 °C caused almost 3-fold increase in the yield coefficient Y_A . It can be hypothesized that it is due to a fast cell growth irrespective of the immobilization. At constant T_{MF} and M_{IC} the OE increase led to decrease in the yield coefficient Y_A (variants 3 and 4).

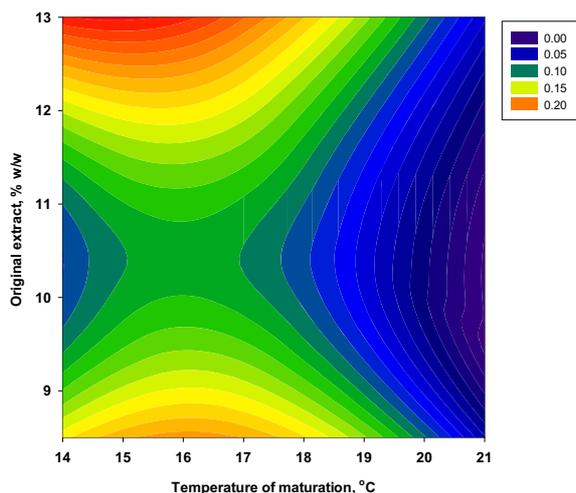


Fig.4. Average VDK reduction rate by biomass in a stationary growth phase ($K_{X,VDK} \cdot X(\text{stat})$)

The specific aldehydes reduction rate ($k_a \cdot A \cdot X$) was relatively constant. The main reason was a phenomenon that was not taken into account by the mathematical model. For all the variants a second peak of aldehydes was observed at the beginning of the maturation. It can be assumed that this was due to the maltotriose uptake. Maltotriose is utilized only in the later stages of alcoholic fermentation, which probably caused new aldehydes synthesis.

The M_{IC} increase and the similar specific aldehydes reduction rate led to an increase in the average aldehydes reduction rate.

The results in Table 2 correspond to the observed fermentation dynamics. The OE increase resulted in longer fermentation time. The increase in M_{IC} and T_{MF} led to a reduction in the primary fermentation time. It should be highlighted that the temperature affected most

significantly the maturation time reduction. On the contrary, the M_{IC} increase was related to the synthesis of more carbonyl compounds, which caused prolonged maturation.

CONCLUSION

A detailed study of the fermentation kinetics and the carbonyl compounds synthesis and reduction in beer fermentation with immobilized cells was carried out. The results showed that the carbonyl compounds kinetics affected significantly on the fermentation time. The carbonyl compounds kinetics was a function of T_{MF} (maturation temperature, respectively), M_{IC} and OE. The results showed that the temperature affected most significantly the carbonyl compounds synthesis and reduction. The M_{IC} increase led to the synthesis of more carbonyl compounds, which caused prolonged maturation. The increase in the wort extract resulted in longer fermentation time.

ACKNOWLEDGEMENTS

We would like to thank Kamenitza Plc for generously supplying us with the wort needed for our experiments.

LIST OF SYMBOLS

- T_{MF} – main fermentation temperature, °C
- OE – original wort gravity, °P
- M_{IC} – immobilized cells mass, g
- $X(t)$ – biomass concentration, g/dm^3 ;
- $P(t)$ – ethanol concentration, g/dm^3 ;
- $S(t)$ – substrate (extract) concentration, g/dm^3 ;
- μ – specific biomass growth rate, h^{-1} ;
- μ_{max} – maximal specific biomass growth rate, h^{-1} ;
- q_p – specific ethanol production rate, $g/(g \cdot h)$;
- q_{pmax} – maximum specific ethanol production rate, $g/(g \cdot h)$;
- K_{SX} – Monod constant for the substrate, g/dm^3 ;
- K_{SP} – Monod constant for the product, g/dm^3 ;
- A – aldehydes concentration, mg/dm^3 ;
- Y_A – yield coefficient for aldehydes, $mg/(g \cdot h)$;
- k_A – reduction coefficient for aldehydes, $mg/(g \cdot h)$;
- VDK – vicinal diketones concentration, mg/dm^3 ;
- Y_{VDK} – yield coefficient for vicinal diketones, $mg/(g \cdot h)$;
- $K_{X,VDK}$ – reduction coefficient for vicinal diketones, $mg/(g \cdot d)$;
- $E(t)$ – error between the experimental and model data;
- R – universal gas constant, $J/(kmol \cdot K)$;
- T – absolute temperature, K ;
- E – activation energy, J/mol ;
- MF – main fermentation;
- MatF – maturation.

REFERENCES

- Analytica EBC. 2004, Fachverlag Hans Carl, Nurnberg.
- Andreas-Toro A., J.M. Giron-Sierra, J.A. Lopez-Orozco, C. Fernandez-Conde, F. Garcia-Ochoa (1998). "A kinetic model for beer production under industrial operational conditions", Mathematics and Computers in Simulation, v. 48, p.65-74

- Branyik, T.; A. Vicente; P. Dostalek; and J.A. Teixeira. 2005 "Continuous beer fermentation using immobilized yeast cell bioreactor systems", *Biotechnol. Prog.*, 21, 653-663
- Briggs, D.; C. Boulton; P. Brookes; and R. Stevens. 2004. *Brewing science and practice*, Woodhead publishing in food and science, CRC Press
- Debourg, A., 1999. "Yeast flavor metabolites", EBC Monograph, 28, p.60–73
- Handbook of brewing: Processes, Technology, Markets. 2009. Hans Michael Esslinger (ed.), ISBN 978-3-527-31674-8, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
- Kostov G.; S. Popova; V. Gochev; P. Kpoprinkova-Hristova; M. Angelov. 2012. "Modeling of batch alcohol fermentation with free and immobilized yeasts *Saccharomyces cerevisiae* 46 EVD." *Biotechnology and Biotechnological equipment*, ISSN 1310-2818, Vol. 26, 3, doi: 10.5504/bbeq.2012.0025
- Marinov M. 2010. "Practice for analysis and control of alcohol beverages and ethanol", *Academic Publisher of UFT*, ISBN 987-954-24-0150-6, p. 196 (in bulgarian)
- Meilgaard, M. 1975. "Flavour chemistry of beer. Part 2: Flavour and threshold of 239 aroma volatiles", *Technical Quarterly of the Master Brewers Association of the Americas*, 12, 151-168.
- Mitev S.V.; S. B. Popova. 1995. "A model of yeast cultivation based on morphophysiological parameters." *J. Chemical and Biochemical Engineering Quarterly*, 3, Zagreb, 119-121.
- Naydenova V, M. Badova, S. Vassilev, V. Iliev, M.Kaneva, G. Kostov, 2014."Encapsulation of brewing yeast in alginate/chitosan matrix: lab-scale optimization of lager beer fermentation", *Biotechnology and Biotechnological Equipment*, eISSN: 1314-3530, accepted, in press
- Parcunev, I, V. Naydenova; G. Kostov; Y. Yanakiev, Zh. Popova; M Kaneva; I. Ignatov, 2012. "Modelling of alcoholic fermentation in brewing – some practical approaches". In: Troitzsch, K. G, Möhring., M. and Lotzmann, U. (Editors), *Proceedings 26th European Conference on Modelling and Simulation*, ISBN: 978-0-9564944-4-3, pp. 434-440
- Popova S. 1997. "Parameter identification of a model of yeast cultivation process with neural network", *Bioprocess and Biosystems Engineering*, 16(4), 243-245, DOI: 10.1007/s004490050315
- Ramirez W.F., Maciejowski, J. 2007. "Optimal beer fermentation" *J. Inst. Brew.* 113(3), 325–333
- Rusell I., 2006, "Yeast", In: Priest F.G. and Stewart G.G (Editors) *Handbook of brewing*, Taylor & Francis Group, LLC, 281-333
- Vassilev S., V. Naydenova, M.Badova, V. Iliev, Maria K., G.Kostov, and S. Popova, 2013. "Modeling Of Alcohol Fermentation In Brewing - Comparative Assessment Of Flavor Profile Of Beers Produced With Free And Immobilized Cells", in: W. Rekdalsbakken, R. T. Bye and H. Zhang (Editors), *Proceedings 27th European Conference on Modelling and Simulation*, ISBN: 978-0-9564944-6-7, p. 415-421.
- Willaert R., 2007." The beer brewing process: wort production and beer fermentation". In: Y.H. Hui (Editor) *Handbook of food products manufacturing*, John Wiley & Sons, Inc., Hoboken, New Jersey, 443-507

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is associated professor at the department "Technology of wine and brewing" at

University of Food Technologies, Plovdiv. He received his MSc in "Mechanical engineering" in 2007 and PhD on "Mechanical engineering in food and flavor industry (Technological equipment in biotechnology industry)" in 2007 from University of Food Technologies, Plovdiv. His research interests are in the area of bioreactors construction, biotechnology, microbial populations investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics.

VESSELA NAYDENOVA is a PhD student at the department "Technology of wine and brewing" at University of Food Technologies, Plovdiv. She received her MSc in "Technology of wine and brewing" in 2005 at University of Food Technologies, Plovdiv. Her research interests are in the area of beer fermentation with free and immobilized cells; yeast specification and fermentation activity. The PhD thesis is named "Possibilities for beer production with immobilized yeast cells"

VASIL ILIEV is a PhD student at the department "Technology of wine and brewing" at University of Food Technologies, Plovdiv. He received his MSc in "Food safety" in 2012 in University of Food Technologies, Plovdiv. His research interests are in the area of immobilized cells reactors modeling and application. The PhD thesis is named "Intensification of the processes in column bioreactor with immobilized cells"

MARIA KANEVA is assistant professor at the department "Technology of wine and brewing" at University of Food Technologies, Plovdiv. Her research interests are in the area of non-alcoholic beverages, herbal extracts for beverages, modeling of extraction processes.

SILVIA POPOVA. Associate Professor Dr. Silviya Popova received her MSc in mathematics from University of Sofia, Bulgaria (1977) and PhD on "New methods for automatization of videomicroscopy microbiological investigation" from the Bulgarian Academy of Sciences (2001). Her research interests are in modeling and identification of biotechnological processes, estimation and control of biotechnological processes, adaptive control, neural networks and image processing.

PETIA KOPRINKOV-HRISTOVA is associate professor at the Institute of Information and Communication Technologies. She received her MSc in "Biotechniques" from Technical University of Sofia in 1989 and PhD on "Processes Automation" from the Bulgarian Academy of Sciences in 2000. Her research interests are in the area of automatic control theory focused on intelligent methods (neural networks and fuzzy systems) with application to nonlinear systems (mainly biotechnological processes).

ROBUST PROCESS CONTROL WITH SATURATED CONTROL INPUT

Frantisek Gazdos and Jiri Marholt

Faculty of Applied Informatics

Tomas Bata University in Zlin

Nam. T. G. Masaryka 5555, 760 01 Zlin, Czech Republic

Email: gazdos@fai.utb.cz

KEYWORDS

Process control, Robust control, Saturated control input, Polynomial approach, Simulation.

ABSTRACT

The contribution presents a methodology to design a robust control loop in case of the saturated control input. The control system design is based on the polynomial approach resulting in the pole-placement problem to be solved. This task is addressed numerically by means of standard MATLAB functions to meet both the constraints on the control input and robustness of the resultant loop. New control quality criteria are suggested for this purpose. The proposed methodology is illustrated practically on a simple simulation example with a classical feedback set-up and one optimized parameter. The presented preliminary results show applicability of the suggested approach.

INTRODUCTION

When designing control systems nowadays it is common to use various simulation tools. It allows to make experiments safely prior to the implementation under real conditions and it saves both time and costs. In some cases, e.g. designing safe control systems for various reactors and unstable processes, it can even save lives when done properly. A simulation model of the process to be controlled is the key crucial point in the designing procedure. It has to contain the main important properties of the process with respect to control. Generally, the more information about the process the better for the control system design. However, a complicated model is not practical for both implementation and control system design as it leads to more complex controllers. Therefore a good process model has to be a trade-off between model complexity and practicability. The fact that only an approximate model of a real nonlinear process is used for the control system design has to be kept in mind and solved using e.g. the adaptive or robust control approach. The adaptive control systems (e.g. Åström and Wittenmark 1995) are generally more complex as they must readjust to new process operating conditions. Robust control systems (e.g. Morari and Zafirov 1989; Barmish 1994; Bhattacharyya et al. 1995) generally use simpler fixed controllers capable of meeting the control requirements not only for one particular process model but for a certain class

of them. Practically the robust controllers are more often used in the industrial practice due to their simpler structure although the achieved control quality may be worse compared to the adaptive control approach.

In practical control applications there are always limits. The most crucial are the constraints on the manipulated variable - the so called control input signal which is used to obtain the desired course of the controlled variable. This signal is always represented by a certain physical quantity, such as a flow rate, electric current or voltage etc which obviously has some limits. Besides amplitude limitations of the manipulated variables there are very often limits on the achievable speed of changes of the variables due to the used actuators, e.g. valves. These facts have to be carefully considered in the control system design procedure and simulation testing. Not respecting these limits can lead to serious consequences, especially when dealing with hardly controllable processes, e.g. unstable, with significant time-delay or with an inverse response (Stein 2003). In the literature there is a great number of classic methods dealing with this problem, often called anti-wind-up techniques applicable mainly to popular PI and PID controllers (e.g. Saberi et al. 2000; Glatfelder and Schaufelberger 2003). Among modern control approaches the predictive control concept is also effective and popular in this field nowadays (Camacho and Bordons 2004; De Doná et al. 2000), although it is more computationally demanding.

Although there are many sources devoted to the robust control systems design and to the constrained control separately, simultaneous solutions of both these problems are still quite rare (e.g. Campo and Morari 1990; Miyamoto and Vinnicombe 1996; Huba 2010). This paper represents a contribution to this interesting and practically important topic. The methodology fruitfully utilized in this contribution is based on the systematic algebraic control concept transforming the control system design problem to the solution of polynomial equations (e.g. Kučera 1993; Hunt 1993; Anderson 1998). After formulation of basic control requirements the polynomial approach enables to find both suitable structure and parameters of controllers. Generally it can lead to more complicated structures of the resultant controllers than the classical PI or PIDs but this does not seem a serious problem nowadays when most of industrial controllers are implemented using PLCs.

A natural part of the procedure for finding a suitable controller using the polynomial approach is the pole-placement problem solution (e.g. Kučera 1994). In this paper this task is solved numerically using the standard MATLAB functions for nonlinear constrained optimization. The resultant poles (free parameters) of the control loop are optimized with respect to both robustness and constraints on the control input signal. For this purpose new control quality criteria and a corresponding procedure are suggested. The whole methodology is illustrated on a simple representative example with the help of simulation means, namely the MATLAB/Simulink environment.

The presented paper is structured as follows: after this introductory section the contribution starts by recalling basics from the algebraic control theory utilized in this work. Next part introduces the control quality criteria for subsequent optimization which is described in detail in the section later. Further parts present the illustrative example, analyse the achieved results and suggest possible areas of future work.

THEORETICAL FRAMEWORK

This section recalls basics of the employed polynomial approach and prepares the space for the methodology respecting both control input limitations and robustness of the resultant loop.

Assume the classical feedback control system of Fig. 1, where G denotes a plant to be controlled by a controller C and the signals w , e , u , and y stand for the reference (set point), control error, control input (manipulated variable), and a process (controlled) variable, respectively. Signals v_u and v_y represent general disturbances.

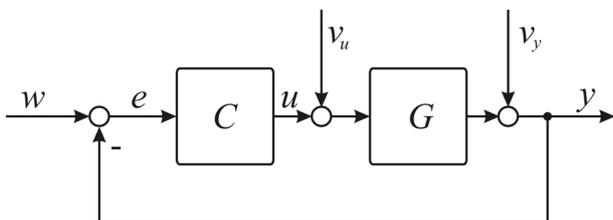


Figure 1: Control System Configuration

Further suppose that in the continuous-time domain both the plant and the controller can be approximated by transfer functions $G(s)$, $C(s)$ with coprime polynomials $b(s)$, $a(s)$ and $q(s)$, $p(s)$ according to (1)-(2) satisfying (3)-(4), i.e. the plant is generally assumed to be strictly proper while the controller is supposed to be proper (the argument s is the complex variable of the Laplace transform).

$$G(s) = \frac{b(s)}{a(s)} \quad (1)$$

$$C(s) = \frac{q(s)}{p(s)} \quad (2)$$

$$\deg a(s) > \deg b(s) \quad (3)$$

$$\deg p(s) \geq \deg q(s) \quad (4)$$

The basic requirements for the control system introduced above are formulated as follows:

- Stability
- Asymptotic tracking of the reference signal
- Disturbances attenuation
- Inner properness

Besides the above-mentioned general requirements, the control system should also be robust to cope with the real nonlinear plant (not only with the adopted linear model) and possible disturbances. In addition the controller has to respect the given physical limitations of its manipulated variable. All these tasks are discussed and solved further in this paper.

From the scheme of Fig. 1 and assuming (1)-(2), it is easy to derive following relationships between the controlled variable $y(t)$ ($Y(s)$ in the complex domain) and the input signals $w(t)$, $v_u(t)$, and $v_y(t)$ ($W(s)$, $V_u(s)$, and $V_y(s)$ similarly); the argument s is omitted in these formulas to keep them more compact and readable:

$$\begin{aligned} Y &= \frac{GC}{1+GC}W + \frac{G}{1+GC}V_u + \frac{1}{1+GC}V_y = \\ &= \frac{bq}{ap+bq}W + \frac{bp}{ap+bq}V_u + \frac{ap}{ap+bq}V_y = \\ &= \frac{bq}{d}W + \frac{bp}{d}V_u + \frac{ap}{d}V_y = \\ &= TW + S_uV_u + SV_y. \end{aligned} \quad (5)$$

Here, d denotes a characteristic polynomial of the closed loop defined as

$$ap + bq = d. \quad (6)$$

Symbols S , T , and S_u represent important transfer functions of the loop known as the sensitivity function, the complementary sensitivity function, and the input sensitivity function, respectively. The sensitivity function S is further used to make the designed control system robust.

Similarly, it is straightforward to derive the formula (7) for the control error $e(t)$ ($E(s)$ in the complex domain):

$$E(s) = \frac{p}{d} [aW(s) - bV_u(s) - aV_y(s)]. \quad (7)$$

Control System Stability

From (5), it is clear that the control system of Fig. 1 will be stable if the characteristic polynomial $d(s)$ given by (6) is stable. This polynomial equation, after a proper choice of the stable polynomial $d(s)$, is used to compute the unknown controller polynomials $q(s)$

and $p(s)$. Roots of the characteristic polynomial $d(s)$ are known as *poles* of the closed loop. Their proper placement influences not only stability of the loop but also the achieved control quality, i.e. a settling-time, overshoots, control input course etc. Therefore the so-called pole-placement problem is a natural part of the polynomial approach to control system design. In this work the poles are optimized numerically to respect both limitations on the control input and robustness of the resultant loop. The procedure is described in detail further in the paper.

Asymptotic Tracking of the Reference Signal and Disturbance Attenuation

Let us assume, as it is a common case, that the reference signal, $w(t)$, can be well approximated by a step function defined in the complex domain as

$$W(s) = \frac{w_0}{s}, \quad (8)$$

for some real w_0 and further, suppose that both disturbances $v_u(t)$ and $v_y(t)$ can be also approximated by the following step functions:

$$V_u(s) = \frac{v_{u0}}{s}, V_y(s) = \frac{v_{y0}}{s}, \quad (9)$$

for some reals v_{u0}, v_{y0} . Then, substituting (8) and (9) into (7) yields

$$E(s) = \frac{p}{d} \left(a \frac{w_0}{s} - b \frac{v_{u0}}{s} - a \frac{v_{y0}}{s} \right). \quad (10)$$

If we suppose $a(0), b(0) \neq 0$ (the common case of proportional systems) then it is obvious that to guarantee zero-control error in the steady state (despite both disturbances), the denominator polynomial of the controller $p(s)$ needs to be divisible by the s term, i.e. the controller has to include an integrator, which will be fulfilled for its denominator polynomial in the following form:

$$p(s) = s \tilde{p}(s). \quad (11)$$

Then the controller (2) can be written as

$$C(s) = \frac{q(s)}{s \tilde{p}(s)}, \quad (12)$$

and the polynomial equation (6) defining stability will be as follows:

$$as\tilde{p} + bq = d. \quad (13)$$

Inner Properness of the Control System

The inner properness of the control system is satisfied if all its parts (transfer functions) are proper. With regard to the strictly proper plant transfer function (1),(3), proper controller (2), (4) and taking into account the solvability of (6) and (13), it is possible to

derive the following formulas for the degrees of the unknown polynomials q, \tilde{p} , and d :

$$\begin{aligned} \deg q(s) &= \deg a(s) \\ \deg \tilde{p}(s) &\geq \deg a(s) - 1 \\ \deg d(s) &\geq 2 \deg a(s). \end{aligned} \quad (14)$$

For practical purposes, when seeking the most simple controller structures fulfilling the given requirements, equalities are taken into account in the inequalities above.

Pole-Placement Problem

For practical computation of the controller's polynomials $q(s), \tilde{p}(s)$ (their coefficients) it is necessary to choose a suitable stable polynomial $d(s)$ appearing on the right side of the polynomial equation (6), (13). This is the so called pole-placement problem mentioned earlier (e.g. Kučera 1994). Therefore we are seeking suitable poles p_i of the designed loop to fulfil the given requirements. Hence the polynomial $d(s)$ can be expressed as:

$$d(s) = \prod_{i=1}^{\deg d} (s - p_i) \quad (15)$$

for some poles (its roots) p_i . Then the control design procedure transforms to the optimization problem of finding the right poles providing the required control quality. When working in the continuous-time domain it is well known the all the poles have to be in the left half of the complex plane to enable stable behaviour, i.e. their real parts have to be negative:

$$\operatorname{Re}[p_i] < 0 \quad \forall i. \quad (16)$$

Besides this it is also well known that complex poles lead to oscillatory behaviour, therefore, it is recommended to employ only real poles for aperiodic (non-oscillatory) behaviour of the system, i.e. it is desirable to have imaginary parts of the poles equal to zero:

$$\operatorname{Im}[p_i] = 0 \quad \forall i. \quad (17)$$

Therefore the optimization problem here can be formulated as to find suitable poles of the loop characteristic polynomial (15) respecting their stability (16) and aperiodic sense (17). The methodology for this task ensuring both loop robustness and limitations on the manipulated controller variable is introduced further in the next section. Based on the information presented above it is suggested to choose the characteristic polynomial of the closed loop $d(s)$ as

$$d(s) = \prod_{i=1}^{\deg d} (s + \alpha_i) \quad (18)$$

for some real constants $\alpha_i > 0$. This ensures both stability of the closed loop (all poles will be negative, i.e. stable, at positions $p_i = -\alpha_i$) and aperiodic behaviour as the poles are real numbers. Now the optimization task is to find optimal values of the free tuning parameters $\alpha_i > 0$.

METHODOLOGY

This section describes the used procedure for optimization of the poles to meet the required control quality, i.e. loop robustness and limitation on the control input. First suitable control quality criteria are suggested and then the methods and procedure of optimization is clarified.

Control Quality Criteria

The solved optimization problem can be formulated as follows:

$$\min_{\alpha} J_{rob}(\alpha) \text{ such that } \alpha_i > 0 \text{ and } J_u(\alpha) = 0, \quad (19)$$

where J_{rob} is the sub-criterion for assessing the loop robustness, J_u the criterion for evaluating demands on the manipulated variable, i.e. control input signal $u(t)$, and α is the vector of optimized parameters α_i . As far as the loop robustness is concerned, a peak gain of the sensitivity function frequency response given by the infinity norm H_{∞} is a good measure for this purpose, e.g. (Skogestad and Postlethwaite 2005). Therefore it is suggested to use the sensitivity function S from (5) and its infinity norm H_{∞} to assess the loop robustness. The sensitivity function is according to (5) given as:

$$S = \frac{1}{1 + GC} = \frac{ap}{ap + bq} = \frac{ap}{d}, \quad (20)$$

then, the first sub-criterion describing robustness reads:

$$J_{rob} = \|S\|_{\infty} = \sup_{\omega} |S(j\omega)|, \quad (21)$$

where ω is the frequency. The second criterion J_u describes the demands on the manipulated variable $u(t)$ and is formed as follows. Let us define the achievable limits of the manipulated variable (control input) as U_{min} and U_{max} where the first one denotes the minimum allowed value of the signal and the latter one the maximum allowed value of the variable. Therefore the control input has to be in the following defined interval:

$$u(t) \in \langle U_{min}; U_{max} \rangle \quad \forall t. \quad (22)$$

Further denote $\Delta u(t)_{max}$ as the maximum overshoot of the manipulated variable above the given limit U_{max} and correspondingly $\Delta u(t)_{min}$ the maximum undershoot of the manipulated variable under the given limit U_{min} . Then the sub-criterion J_u is computed according to this simple formula:

$$J_u = \Delta u(t)_{max} + \Delta u(t)_{min}. \quad (23)$$

It is evident that the sub-criterion is equal to zero if the manipulated variable is within the desired limits and it is positive with higher values for control input out of the required range. The situation is well illustrated in the following picture, Fig. 2, with the limits chosen as $U_{min} = -1$ and $U_{max} = 1$.

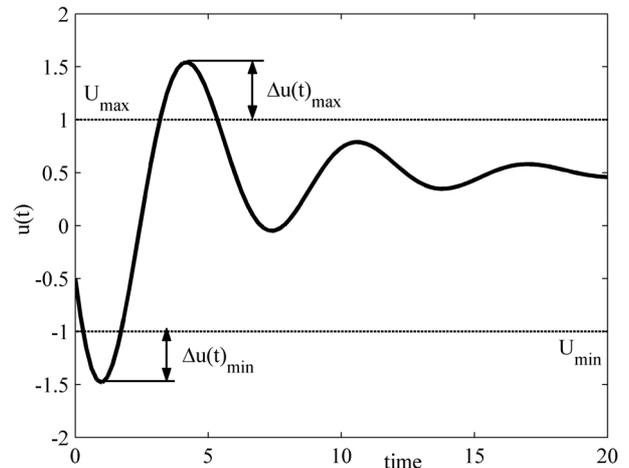


Figure 2: Explanation of the Sub-Criterion J_u

It is obvious that achieved values of the control quality criteria in (19) explained above depends on the placement of the closed-loop poles in (15), i.e. on the choice of the free tuning parameters $\alpha_i > 0$ in (18). This is solved with the help of simulation means using the procedure described in the next section.

Optimization Procedure

First, let us scale the loop variables so that both the controlled output $y(t)$ and correspondingly also the reference signal $w(t)$ are within the range from zero to one, i.e.

$$y(t), w(t) \in \langle 0; 1 \rangle. \quad (24)$$

Then the worst-case behaviour of the control system (regarding the changes in the reference signal) can be analysed by considering the reference change of magnitude one. Therefore the designed control system is analysed facing this condition and the control quality criteria in (19) are assessed for different values of the free tuning parameters $\alpha_i > 0$. This is done with the help of simulation means, namely the MATLAB environment and its toolboxes for simulation and optimization. The procedure can be briefly described as follows:

1. choose number of optimized parameters α_i (the closed-loop poles $p_i = -\alpha_i$ can be different or multiple);
2. for every α_i choose an interval for optimization;
3. find a solution of the problem specified in (19), i.e. find a minimum of the sub-criterion J_{rob} subject to

the conditions such that $J_u = 0$ on the given region of α_i from the previous point;

4. record the resultant parameters α_i and verify if they fulfil the given requirements.

If the algorithm for solution of the problem fails then there may not be such combination of α_i (under the given conditions) to respect the limits of the control input $u(t)$. Then the designer has several basic possibilities: try to increase the number of optimized parameters α_i , use different control structures (e.g. the 2DoF control set-up with a pre-filter of the reference signal), or has finally no other way than enlarging the prescribed limits on the manipulated variable $u(t)$.

The algorithm for the optimization uses a standard MATLAB function for nonlinear constrained minimization (nonlinear programming) *fmincon*. It is a gradient-based method — the trust-region-reflective algorithm based on the interior-reflective Newton method, described in detail in, e.g. (Coleman and Li 1996). For this algorithm it is necessary to choose also a starting point (initial estimate) of the optimization. With no prior information, this is usually the middle of the interval from the point 2 of the optimization procedure described above.

Next section illustrates the suggested methodology using a simulation study — robust control system design with a saturated input signal applied to a given representative simulation model.

ILLUSTRATIVE EXAMPLE

Consider a system described by the following transfer function in the continuous-time domain:

$$G(s) = \frac{b(s)}{a(s)} = \frac{18400}{s^2 - 2.418s - 3998}. \quad (25)$$

This transfer function approximates behaviour of the magnetic levitation system CE 152, the product of TQ Education and Training Ltd designed for studying system dynamics and experimenting with control algorithms. The system consists of a coil levitating a steel ball in the magnetic field with the position sensed by an inductive linear sensor connected to an A/D converter. The coil is driven by a power amplifier connected to a D/A converter. A basic control task is to control the position of the ball freely levitating in the magnetic field of the coil. From the control theory point of view, the magnetic levitation system is a nonlinear unstable system with one input and one output and relatively fast dynamics. Detailed description of the apparatus can be found in e.g. (Gazdoš et al. 2009). The model (25) represents a linear approximation of the system levitating the ball in the middle of the space. Following the polynomial approach methodology described in the previous sections a suitable controller for this system is designed in the following general form:

$$C(s) = \frac{q(s)}{p(s)} = \frac{q(s)}{s\tilde{p}(s)} = \frac{q_2s^2 + q_1s + q_0}{s(\tilde{p}_1s + \tilde{p}_0)}, \quad (26)$$

hence, it is a real (filtered) PID controller. Unknown coefficients of the controller are obtained by the solution of the polynomial equation (13) for a given stable characteristic polynomial $d(s)$. This polynomial, in the general form (15),(18), must be according to (14) of the 4th degree and it is suggested in the following simple form:

$$d(s) = (s + \alpha)^4, \quad (27)$$

for some real parameter $\alpha > 0$ subject to optimization. Then the closed-loop has 4 identical poles located at $p_{1,2,3,4} = -\alpha$ which will guarantee both stable and aperiodic behaviour. Although this simple choice limits possibilities of achievable control quality it enables simple tuning of the loop and it is used here to illustrate simply the methodology introduced in this paper. Now the one free tuning parameter α is optimized numerically according to the procedure suggested in the previous section to respect both robustness of the loop and limitations on the control input signal $u(t)$ which are in this case defined as:

$$u(t) \in \langle -1; 1 \rangle \quad \forall t. \quad (28)$$

In this simple case of only one tuning parameter it is possible to obtain easily the course of the sub-criterion J_u (23) assessing the control input signal with respect to the given limitations (28), depending on the parameter α . It is recorded in the following picture, Fig. 3.

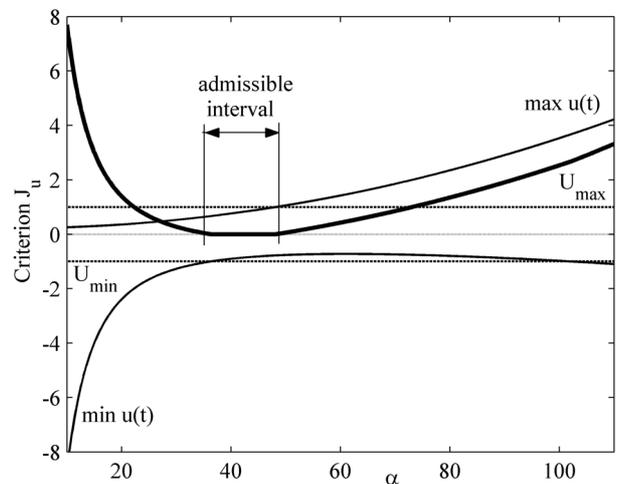


Figure 3: Sub-Criterion J_u with α

From the plot, it is obvious that there exists an interval where $J_u = 0$, i.e. there are values of α for which the given control input limitations are respected. Further inspection of the results shows that this interval is for $\alpha \in \langle 36.4; 48.1 \rangle$ approximately. If we choose several values of the parameter from and outside of the admissible interval to test the results, it is possible to obtain the recorded simulation of the control input $u(t)$ as presented in the next figure, Fig. 4.

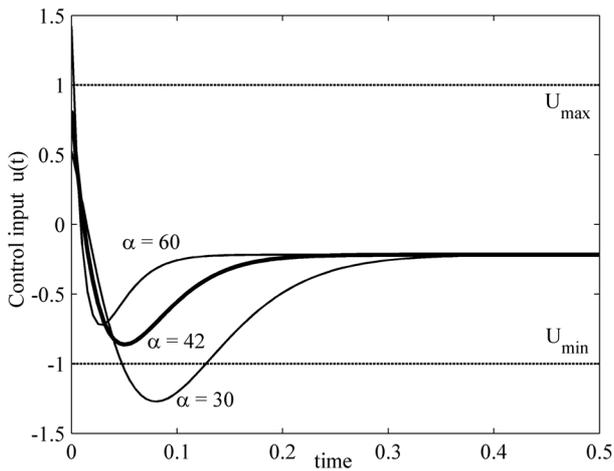


Figure 4: Control Input $u(t)$ with α

From the graph, it is clear that for α from the suggested interval ($\alpha = 42$) the control input is in the prescribed limits while for $\alpha = 30$ and $\alpha = 60$ (outside the interval) it is out of the required range $\langle -1; 1 \rangle$.

Next figure, Fig. 5, shows the course of the sub-criterion J_{rob} (21) assessing the loop robustness on the admissible interval of the parameter α .

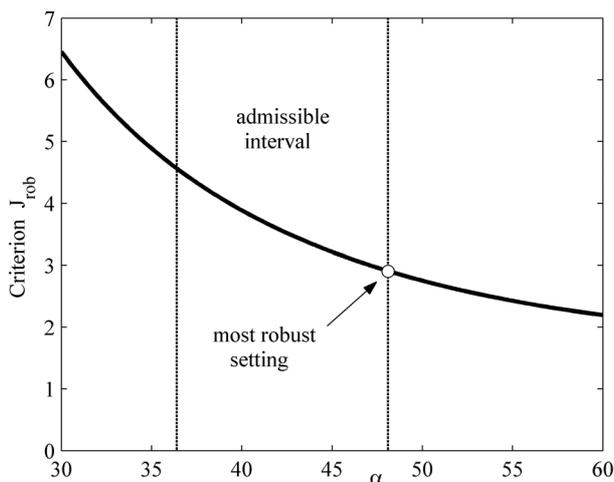


Figure 5: Sub-Criterion J_{rob} with α

The figure reveals that the most robust setting respecting the given control input limits is on the right end of the admissible interval. Therefore it is for $\alpha = 48.1$ approximately. For safety reasons the *optimal* value of this parameter is suggested a bit smaller, for $\alpha = 48$. This value respects the control input limitations and leads to the most robust setting of the designed control system, under given conditions (control set-up, number of optimized closed-loop poles, given limits on the control input signal). Simulated responses of the controlled variable and control input signal are presented in the next graphs, Fig. 6 and Fig. 7. In these experiments step-disturbances of 10% amplitude were injected at times $t = 1$ (acting on the control input $u(t)$) and $t = 3$

(acting on the controlled variable $y(t)$) to test robustness of the designed loop. Two settings of the optimized parameter are recorded to assess the robustness: $\alpha = 48$ (suggested optimal setting - most robust) and $\alpha = 37$ (less robust setting, see Fig. 5).

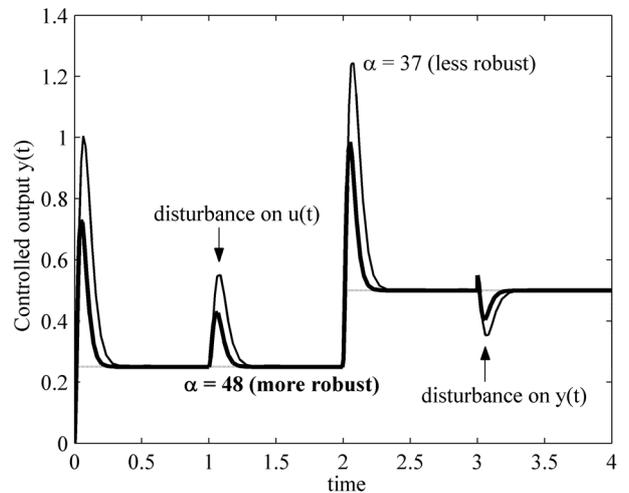


Figure 6: Controlled Output Response

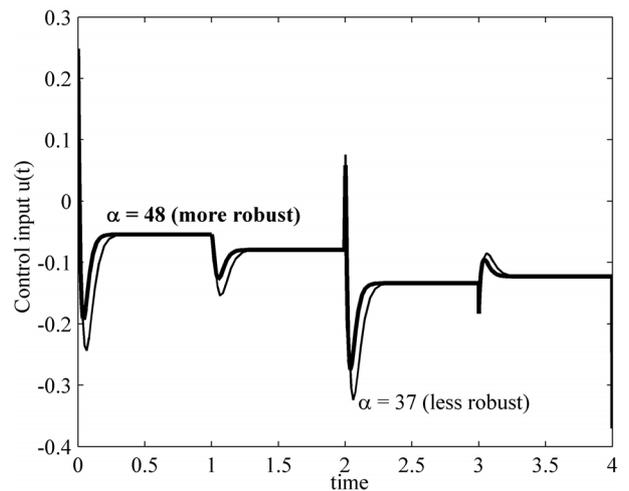


Figure 7: Control Input Response

The graphs show that the controlled variable tracks its desired value stably without a steady-state control error. Although there are relatively high overshoots the response is aperiodic, both disturbances are compensated and the control input is in the prescribed limit $\langle -1; 1 \rangle$. When comparing the robustness of the designed loop for the suggested *optimal* setting ($\alpha = 48$) to another one, not so robust ($\alpha = 37$) it is obvious that the recommended *optimal* setting provides better results - with smaller overshoots, faster settling-time, better compensation of both disturbances and less demands on the manipulated variable.

Concerning the higher amplitude of overshoots in the presented control responses - it is attributable to the fact that i) the system is unstable (such systems are

more difficult to control), ii) most robust setting of the loop respecting limits on the control input is desired (this results in responses of worse control quality with bigger overshoots and longer settling-times), iii) the classical feedback control set-up was employed (another control configurations with, e.g. a filter of the reference signal can lead to comparably smaller overshoots). Therefore for future work is recommended to use e.g. the 2DoF (2 degrees-of-freedom) control set-up with one feedback and one feedforward controller filtering the reference signal. This will decrease the overshoots and enables to achieve better control quality while respecting the limits on the manipulated variable. It is also expected that higher number of optimized parameters α_i (consequently, also closed-loop poles p_i) enables to achieve better results for both control quality and robustness.

CONCLUSIONS

This paper investigates and presents one possible way how to design a robust control system in the presence of constraints on the control input signal. For this purpose the systematic polynomial approach is fruitfully exploited resulting in a solution of polynomial equations. Robustness and input constraints are solved numerically using standard MATLAB functions for nonlinear constrained optimization and new suggested criteria. Optimal poles of the designed loop are the results of this optimization procedure which is based on the usage of simulation means. Presented preliminary results show the potential of the suggested methodology even though the illustrated example is simple with only one optimized parameter. Better results are expected in case of optimizing more parameters and with usage of different control configuration which is the course of future works. An extension to cover not only control input amplitude limitations but also constraints on the speed of changes of this signal is relatively simple. Usage of the suggested procedure for MIMO (Multi-Input Multi-Output) systems is also open.

REFERENCES

Anderson, B.D.O. 1998. "From Youla-Kucera to Identification, Adaptive and Nonlinear Control." *Automatica*, Vol.34, 1485-1506.

Åström, K. J. and B. Wittenmark. 1995. *Adaptive Control*. Addison-Wesley, Reading, Massachusetts.

Barmish, B.R. 1994. *New Tools for Robustness of Linear Systems*. Macmillan.

Bhattacharyya, S.P.; H. Chapellat and L. H. Keel. 1995. *Robust Control—The Parametric Approach*. Prentice-Hall.

Camacho, E.F. and C. Bordons. 2004. *Model Predictive Control*. Springer-Verlag, London.

Campo, P.J. and M. Morari. 1990. "Robust Control of Processes Subject to Saturation Nonlinearities." *Computers & Chemical Engineering*, Vol.14, No.45, 343-358.

Coleman, T.F. and Y. Li. 1996. "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds." *SIAM Journal on Optimization*, Vol.6, 418-445.

De Doná, J.A.; G.C. Goodwin and M.M. Seron. 2000. "Antiwindup and Model Predictive Control: Reflections and Connections." *European Journal of Control*, Vol.6, No.5, 467-477.

Gazdos, F.; P. Dostal and R. Pelikan. 2009. "Polynomial Approach to Control System Design for a Magnetic Levitation System." *Cybernetic Letters*, 1-19.

Glattfelder, A.H. and W. Schaufelberger. 2003. *Control Systems with Input and Output Constraints*. Springer, London.

Huba, M. 2010. "Robust Constrained PID Control". In *Proceedings of the International Conference Cybernetics and Informatics* (Vyšná boca, Slovak Republic, Feb.10-13). 1-18.

Hunt, K.J. 1993. *Polynomial Methods in Optimal Control and Filtering*. Peter Peregrinus Ltd., London.

Kučera, V. 1993. "Diophantine Equations in Control—a Survey." *Automatica*, Vol.29, 1361-1375.

Kučera, V. 1994. "The pole placement equation. A survey" *Kybernetika*, Vol.30, No.6, 578-584.

Miyamoto, S. and G. Vinnicombe. 1996. "Robust Control of Plants with Saturation Nonlinearity based on Coprime Factor Representations". In *Proceedings of the 35th IEEE Conference on Decision and Control* (Kobe, Japan, Dec.11-13). IEEE, 2838-2840.

Morari, M. and E. Zafirov. 1989. *Robust Process Control*. Prentice Hall, New Jersey.

Saberi, A.; A.A. Stoorvogel and P. Sannuti. 2000. *Control of Linear Systems with Regulation and Input Constraints*. Springer, London.

Skogestad, S. and I. Postlethwaite. 2005. *Multivariable Feedback Control: Analysis and Design*. Wiley, Chichester.

Stein, G. 2003. "Respect the Unstable." *IEEE Control Systems Magazine*, Vol.23, No.4, 12-25.

AUTHOR BIOGRAPHIES



FRANTIŠEK GAZDOŠ was born in Zlín, Czech Republic, in 1976, and graduated from the Brno University of Technology in 1999 with an MSc degree in Automation. He then followed studies of Technical cybernetics at Tomas Bata University in Zlín, obtaining a PhD degree in 2004. He became an

associate professor for machine and process control in 2012 and now works at the Department of Process Control, Faculty of Applied Informatics, Tomas Bata University in Zlín, Czech Republic.

He is an author or a coauthor of more than 70 journal contributions and conference papers giving lectures at foreign universities, such as Politecnico di Milano, University of Strathclyde Glasgow, and Universidade Técnica de Lisboa, among others. His research activities cover the area of modelling, simulation, and control of technological processes. E-mail: gazdos@fai.utb.cz.



JÍŘÍ MARHOLT was born in Zlín, Czech Republic in, 1983, and graduated from the Tomas Bata University in Zlín in 2008 with an MSc. degree in Automation. He is now finishing his Ph.D. studies of Technical Cybernetics at the Department of Process Control, Faculty of Applied Informatics of

Tomas Bata University in Zlín, Czech Republic. His dissertation thesis is focused on the control of unstable systems. Besides this, his research interests cover also the area of process modelling, simulation and algebraic approach to control system design.

COMPUTATIONAL MODEL FOR SPRAY QUENCHING OF A HEAVY FORGING

Mahdi Soltani
Annalisa Pola
Giovina Marina La Vecchia
Dipartimento di Ingegneria Meccanica e Industriale (DIMI)
Università degli Studi di Brescia
Brescia, Italy
mahdi.soltani@unibs.it
annalisa.pola@unibs.it
marina.lavecchia@unibs.it

KEYWORDS

Spray Quenching; Simulation; Modelling; Residual Stress

ABSTRACT

Considering heavy forged parts presented that water spray is an appropriate technique as quenching operation. In general, the rapid cooling causes residual stress field due to the dis-homogeneity of the distribution of temperature in the part and the different microstructural phases. In comparison with other conventional quenching methods, spray-quenching method could have a capability to control the temperature distribution through the surface. In order to gain entry the appropriate mechanical properties, the spray control valves need to adjust to the foundation of heavy forgings cross sections. In this situation, the modelling of the spray quenching operation necessarily deserves for more elaboration. The present paper is aimed at developing and simulating a spray quenching process of a heavy forged shaft produced in a hardening and tempering steel.

INTRODUCTION

The heat transfer of convection dependency on the quenchant velocity leads to establishing different methods for cooling such as film quenching, immersion quenching and spray quenching. At immersion quenching process, stationary fluid has been progressively switched to a flow with special velocity depending on vapour flow; in addition, this fluid flow can superimpose on a period of enforced bath flow. Due to working on a large scale of heavy forgings with different cross sections, there is always a possibility to create a massive non-homogeneous thermal distribution, non-uniform microstructural transformations and eventually unwanted residual stress (Schajer, 2013). Spray quenching technique has a capability to adapt sprays of water upon different sectors of a complex shaped heavy forging, in order to achieve demanding almost homogenous metallurgical properties in the different cross sections (Hodgson, et al., 1968). Therefore, after a quenching treatment, a steel forged part should be characterized by the presence of Martensite at prefixed depths from the surface, to

achieve proper hardness, tensile stress and impact resistance of the final tempered component, in agreement with specific standards.

There are also some technical concerns about spray cooling, for instance the lack of repeatability of quenching process for apparently identical nozzles; that it can be modified by utilizing corrosion resistant nozzles and adopting stringent quality control and nozzle characterization practices (Hall & Mudawar, 1995). Furthermore, spray quenching has been introduced as one of methods developing more efficient production to remarkably diminishing cost.

Not only does spray quenching method need to be technically developed, but it also needs to be evaluated the spray quenching formula describing the connection between local heat flux and surface temperature. It can be fully functional to determine a sufficient justification for the dedicated study and evaluate the effect of various spray quenching parameters.

In general, the heat flux during a quenching treatment follows the boiling curve regimes, as shown in Figure 1. The high temperature heat flux region of this curve corresponds to the film boiling process, generally attributed to the growth of a vapour blanket covering stably the surface as a thermal insulation. Gradually, after the surface temperature reaches the minimum (Leidenfrost point), the collapsing of the bubble vapour starts and hence heat flux radically increase (transition boiling regime) to reaching maximum one.

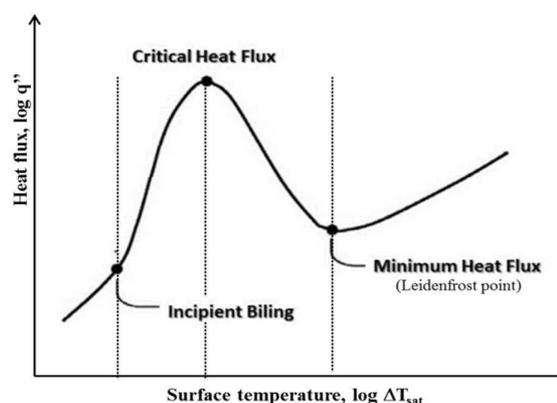


Figure 1: Boiling Curve

Subsequent to critical heat flux, the part would experience sequentially nucleate boiling regime and lastly single-phase regime when the heat flux is significantly falling down. Therefore, the surface will cool down rapidly prior to reach the single phase boiling (Vorster, et al., 2009).

The lack of adequate understanding about spray quenching led to the pertinent literature being quite sparse. Moreover, in comparison with other cooling system, wide variety of parameters (water pressure, nozzle-part distance, etc.) involved on proposed spray quenching constitutive equations causes to introduce complex quenching models able to reliably describe the heat flux through the surface of the metallic part (Rybicki & Mudawar, 2006) (Mascarenhas & Mudawar, 2010) (Wendelstorf, et al., 2008).

The aim of this paper is to validate one of the proposed models for flat parts and starting to expand it on a real industrial cylindrical heavy forging to take into account the massive heat flux, especially at the core of the largest diameter of experimented shaft.

In order to optimize the spray quenching results, it has been demanded the implementation of an intelligent spray system with special identification of behaviour for air and water supply conditions considered to quench a component as rapidly and uniformly to achieve particularly appropriate microstructure and mechanical properties.

DETERMINING HEAT TRANSFER COEFFICIENT FOR SPRAY WATER COOLING OVER PLATE

This study is based on using the spray quenching constitutive equations proposed by Wendelstorf, et al. (2008) to determine heat transfer coefficient. This method is a connection between the temperatures of the surface (in this research work assumed as a plate), the fluid layer thickness (d_f) and the water impact density (V_s). Furthermore, the second parameter (water impact density) is mentioned as very important factor having effect on determining of the heat transfer between water spray and plate.

The heat transfer coefficient (HTC) h is defined by:

$$h = q / (T_s - T_w) \quad (1)$$

where T_s and T_w are the temperature of the surface and of the water respectively, and the heat flux density (q) itself run through several characteristic regimes. The heat is transferred through natural convection until boiling occurs by the formation of isolated bubbles.

The heat transfer Constitutive Equations proposed Wendelstorf, et al. (2008) take into account the influence of the impact density, defined as follows:

$$V_s \pi r_{\max}^2 = 2 \pi r_{\max} \cdot d_f \cdot v_f \cdot \rho_f \quad (2)$$

where r_{\max} is the radius of the water spray impact area, v_f is the outward velocity developed by the water film and ρ_f is the density of the film fluid.

Therefore, the coefficient is:

$$h(\Delta T, V_s) = 190 \pm 25 + \tanh\left(\frac{V_s}{8}\right) \cdot \left(140 \pm 4 \cdot V_s \cdot \left[1 - \frac{V_s \cdot \Delta T}{72000 \pm 3500}\right] + 3.26 \pm 0.16 \cdot \Delta T^2 \cdot \left[1 - \tanh\left(\frac{\Delta T}{128 \pm 1.6}\right)\right]\right) \quad (3)$$

SPRAY QUENCHING OPERATION OVER HEAVY FORGING

Traditional immersion quenching process does not provide a substantial control of cooling; this is principally identified with the final mechanical property of heavy forging part. In fact, rectification is not able to apply right through the quenching operation thus it is not possible to control the heat flux density dissimilitude.

In comparison to other quenching techniques, spray quenching could have capability to be sufficient to cover everything that needs doing. The quenching technical competence of the spraying processes is contingent upon the typical spray and hence the water pressure plays an important role during operation (Liscic, et al., 2010). Spray quenching has ability to vary water pressures and mass flow, being the spray system equipped with servo valves to supply a continual difference of droplets velocity. Spray quenching with lower mean drop velocity obviously would suffer reduced residual stress, however the appropriate mechanical properties would not be eventually achieved (Hossain, et al., 2006).

In this situation, it is absolutely essential to confirm the continuousness of the quenching operation in reaction to water pressure variations; therefore, the time of working spray valves should be proportionate to the gravity of diameter of cylindrical heavy forging shaft.

One of the most outstanding features of spray quenching operation is how to control cooling process by making a sufficient adjustment through the supply pressure/mass flow ranges. Therefore, the operation needs to have a timetable to determine process parameters for each part of a complex shaped heavy forging under spray cooling. It can raise the possibility to reach the quenching term between air and cold water cooling.

Pola et al. (2013), for instance, pointed out a model for spray quenching operation through a heavy forging part; they employed equations proposed by Wendelstorf (2008) to calculate the heat transfer coefficient over the case of simple non-uniform shaped forging. The spray quenching method is based on the assumption that in each impingement area an homogenous distribution of water is considered. As a result, a good agreement was obtained between measured and predicted temperatures at internal locations, even those placed next to the core.

The purpose of this paper is to modify the method proposed by Pola et al. (2013) for a real enormous shaft 14m long and characterized by different diameters (as the largest diameter is 1.22m), i.e. different distances from the nozzles. Because of working on a real huge part, the analysis stopped after 5 hours when the surface reaches almost the temperature of water (20°C) and in

addition, the local temperature loses significantly at the core of the largest diameter.

The method, acquired by Pola et al. (2013), introduced only one part of the pilot plant under cycling quenching; nevertheless, water pressure and total water flow kept constant in other parts.

The present study is aimed at applying a highly complex quenching process with different cycling quenching stages being dependent locally on shaft diameters, to create adequate temperature distributions. Thus, the current case study demands to implement an intelligent system to control the water pressures and mass flow with respect to the different diameters shaft.

Hence, the cooling system has been divided into 5 different regions, as shown in Figure 2. In each nozzles sector the spray parameters were adjusted to reduce or increase the cooling, in order to achieve a uniform treatment.

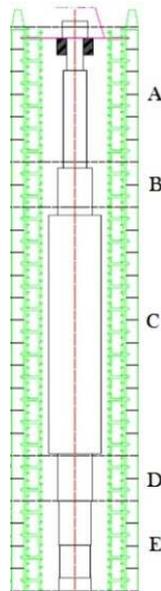


Figure 2: Heavy Forging Shaft

According to the spray characteristics presented in the Table 1, the amount of heat transfer coefficient can be determined through the different diameters by utilizing the constitutive equation proposed by Wendelstorf, et al. (2008). Figure 3 shows the heat transfer coefficient obtained for the largest shaft diameter (1220 mm).

Table 1: Spray Characteristics Used for the Steel Shaft Examined (26NiCrMoV115) under the Pressure 3 bar

Spray angle θ ($^{\circ}$)	60
Temperature of Water ($^{\circ}$ C)	20
Total flow rate $Q \times 10^6$ ($m^3/s m^2$)	9.73E-05
Mean drop velocity v_f (m/s)	15.75
water impact density V_s (kg/s)	6.35
Average Rotation of Shaft (rpm)	3

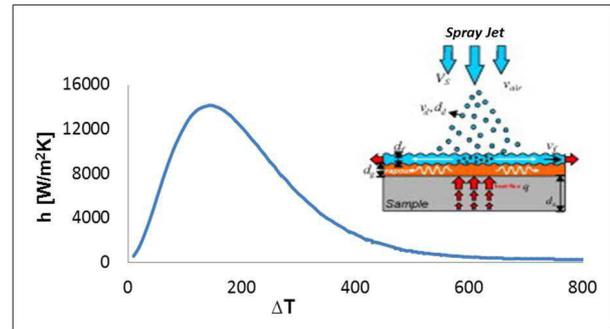


Figure 3: Heat transfer Coefficient through the Largest Shaft Diameter

Spray cooling simulations provide a process control with several vital clues and predict the distribution of the properties over heavy forgings.

The simulation process is composed of five stages, as highlighted in Figure 4. The function model carries on being established on the 3D extruded model.



Figure 4: Modelling Process

The used model was composed of two general subdivisions, a solid shaft and the water spray impingement region. The spray impingement, as shown in Figure 5, is divided into two direct primary spray and secondary spray modelling domains. Regardless of overlapping spray impingement, the area under direct spray impingement would experience much more heat flux; therefore, the corresponding area in the model is assumed with higher heat transfer coefficient (primary part) rather than the area located far away from the direct effect of spray (secondary one). Initially, the feature and 3D modelling is individually extruded utilizing 3D CAD Design Software Solid-Works package.

Afterwards the process was modelled by using the finite element software ProCAST (a trademark of ESI Group). It has the capability of mesh generation, defining material properties and boundary condition, calculating of temperature as a time function based on inputting heat transfer coefficient data.

The temperature dependent thermal properties of the steel (conductivity, density, specific heat) calculated by means of Computherm Database[®] (Pan Iron 5.0) available in Procast, as a function of steel chemical composition and using the “Back Diffusion” model.

The interface between direct water spray and forging surfaces was modelled using the equation proposed by Wendelstorf et al. (2008) (see Fig. 3 for the primary cooling coefficient). For secondary cooling a corrective coefficient to the Wendelstorf equation was used (Pola et al. 2013). The spray cooling process should be controlled

by working spray valves; when they are blocked, the component would undergo only air cooling.

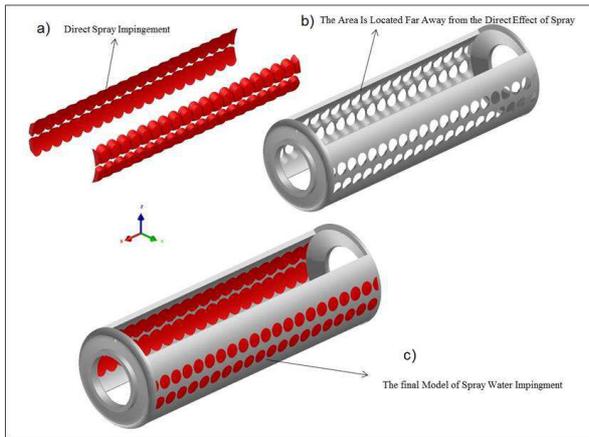


Figure 5: Modelling of Spray Impingement, Direct Spray Impingement (a), Secondary Water Spray Modelling (b) and Final Modelling (c)

The boundary conditions need to be created according to the spray characteristics used for the shaft examined, furthermore, the effect of air cooling outside. Therefore, a value of $15\text{W/m}^2\text{K}$ at 20°C of the heat exchange coefficient was imposed on both the forging surface in contact with air and the external surface of water domains. The initial temperatures fixed at 860°C for the component and 20°C for the water. A rotational speed of the part imposed according to the industrial procedure (3rpm, Fig. 6). Finally; post processing and outcome analysing is done by using ESI Visual Environment package.



Figure 6: Final Spray Quenching Model Highlighted Primary and Secondary Parts

QUENCH CURVE RESULTS AND DISCUSSION

The experimental shaft would be a symmetric shape thus the half-cell displayed in the following figures. As mentioned, ProCAST software package is utilized for transient analysis through this case. The post-processing step is as the output results for analysing and observing the determined plotting or lines through the temperature-time curve.

To evaluate the effectiveness of the cooling the larger diameter considered, taking into account 16 points equally distributed from the centre to the core along the radius distanced 4 cm each other.

Contemplating upon spray cooling of a heavy forged part in different recorded time that radial temperature is inhomogeneous along the diameter, from the core to surface (Figure 7).

As expected, when the operation starts (i.e. after just 10 min), there is a clear distinction between the temperature of surface and the core; in general, the temperature of the part stays constant at 860°C , except the surface temperature falling radically to the water spray temperature. The temperature of core decreases and, after 2 hours, it reaches to 650°C . Finally, after five of hours spray quenching operation, the temperature of core would fall to 200°C .

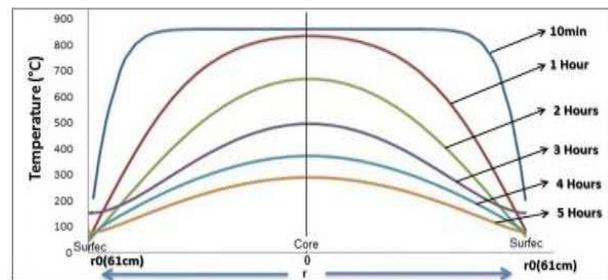


Figure 7: Radial Temperature Distributions in Different Recorded Times

In this situation, at the commencement of the spray quenching, cylindrical shaft surface shrinks more quickly rather than the core, therefore, the surface domain is prone to tensile stresses longitudinally and radial compressive one tangentially. In a hypothetical situation behaving linear-elastic, the deformation achieves equilibrium, thus the stresses need to balance out along opposite direction in the core. Thermal shrinking conduct has profound implications for leading to various local and temporal strains, and thermal residual stresses (Schajer, 2013).

In order to predict how to achieve demanding metallurgical properties after a spray quenching operation through the different sectors of the heavy forging part, the cooling curves obtained by the simulation can superimpose to the continuous cooling transformation curve for the 26NiCrMoV115 steel.

The CCT curves used to assess the microstructure obtained with the quenching procedure was calculated by means of the JMatPro[®] software, assuming an austenitization temperature of 860°C and an ASTM austenitic grain size number of 6.

Figure 8 illustrates that, at any specific time, the different locations could experience distinct cooling conditions. In particular, the cooling curves related at 3 points on the middle sections of the larger and the smaller diameter were taken into account and compared to analyse the quenching effectiveness. The three points are located at the surface, at a quarter and at a half of the

radius that are zones where the mechanical specimens can be requested by the user, depending from the forged geometry and the standards used to control the part at the end of the production.

It can observe that a mainly tempered martensitic microstructure can obtain close to the surface, where the cooling rate is the highest, in both the larger and smaller diameter. On the other hand, the cooling trajectories of inner points intersect high temperature transformations curves and the resulting microstructure contains also large amounts of Bainite. Concerning the centre of the forging part, the cooling rate is so low that almost ferrite and perlite can be obtained. This was confirmed by the microstructural analyses performed on the real part, as shown for instance in Figure 9, in the case of the smaller section and at a half of the radius.

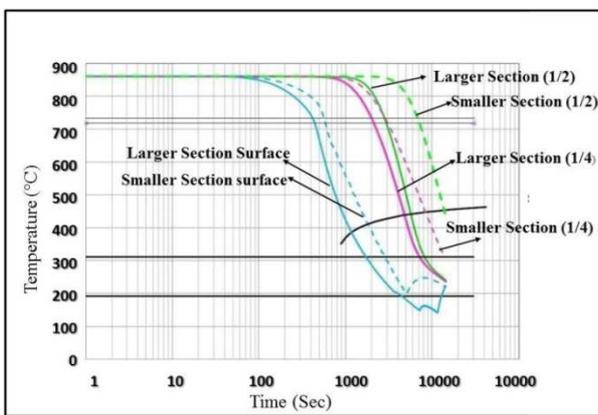


Figure 8: Cooling Curves for the 26NiCrMoV115 Heavy Forging under Pressure of 3 bar

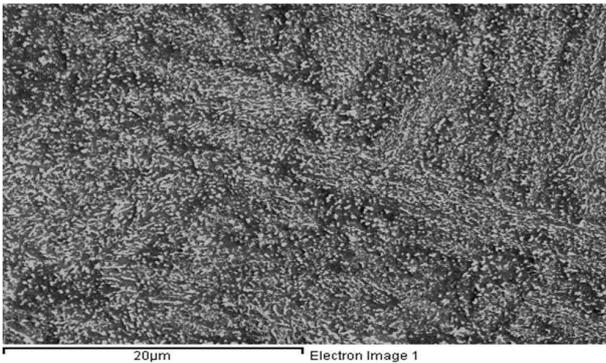


Figure 9: Tempered Martensite at a half of the radius in the smaller section of the forging

CONCLUSION

The present purpose was of developing a CAD based system to model and simulate a spray quenching process. In this case, the part is cooled under a package of full cone spray nozzles with constant pressure. This

study was achieved by using an intelligent system modelling during spray quenching operation for a heavy forging part. The simulation allowed demonstrating the effectiveness of the spray quenching method and the opportunity to predict the final microstructure at the different cross sections.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Mr. F. Zola and MSc. Eng. A. Ferrari (Ofar S.p.A. – Visano) for their technical support and for the permission to publish the test results.

REFERENCES

- Hall, D. & Mudawar, I., 1995. “Predicting the impact of quenching on mechanical properties of complex-shaped aluminum alloy parts” *Heat Transfer in Manufacturing*, 117(2), pp. 479-488.
- Hodgson, J., Saterbak, R. & Sunderland, J., 1968. “An Experimental Investigation of Heat Transfer From a Spray Cooled Isothermal Cylinder” *J. Heat Transfer*, p. 457-463.
- Hossain, S., Truman, C., Smith, D. & Daymond, M., 2006. “Application of quenching to create highly triaxial residual stresses in type 316H stainless steels” *International Journal of Mechanical Sciences*, 48(3), p. 235-243.
- Liscic, B., Tensi, H., Canale, L. & Totten, G., 2010. *Quenching Theory and Technology*. 2nd ed. s.l.:Taylor and Francis Group .
- Mascarenhas, N. & Mudawar, I., 2010. “Analytical and computational methodology for modeling spray quenching of solid alloy cylinders” *International Journal of Heat and Mass Transfer*, 53(25-26), pp. 5871-5883
- Pola, A., Gelfi, M. & Vecchia G.M La, 2013. “Simulation and validation of spray quenching applied to heavy forgings” *Journal of Materials Processing Technology*, 213(12), p. 2247-2253.
- Rybicki, J. R. & Mudawar, I., 2006. “Single-phase and two-phase cooling characteristics of upward-facing and downward-facing sprays” *International Journal of Heat and Mass Transfe*, 49(1-2), pp. 5-16.
- Schajer, G., 2013. *Practical Residual Stress Measurement Methods*. s.l.:A John Wiley & Sons, Ltd..
- Vorster, W., Schwindt, S., Schupp, J. & Korsunsky, A., 2009. “Nozzle, Analysis of the spray field development on a vertical surface during water spray-quenching using a flat spray” *Applied Thermal Engineering*, 29(7), pp. 1406-1416.
- Wendelstorf, J., Spitzer, K.-H. & Wendelstorf, R., 2008. “Spray water cooling heat transfer at high temperatures and liquid mass fluxes”. *International Journal of Heat and Mass Transfer*, 51(19-20), pp. 4902-4910.

MODELLING AND SIMULATION OF WATER TANK

Jiri Vojtesek, Petr Dostal and Martin Maslan
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nam. TGM 5555, 760 01 Zlin, Czech Republic
E-mail: {vojtesek,dostalp}@fai.utb.cz

KEYWORDS

Modelling, Simulation, Mathematical Model, Numerical Methods, Water Tank.

ABSTRACT

The modelling and simulation play a very important role in the industry where it can help with the description of the system and the choice of the optimal control strategy. This contribution is focused on the modelling and simulation procedure which usually precedes the design of the controller. The mathematical model is derived with the use of material balance and produces nonlinear Ordinary Differential Equation (ODE). The static analysis provides optimal working point and the dynamic analysis gives an overview about the behavior of the system. Mentioned procedure is tested on the real model of the water tank as a part of the process control teaching system PCT40 from Armfield. Results have shown that proposed mathematical model is accurate and can be used for the design of the appropriate controller.

INTRODUCTION

The modelling and simulation are important tools often used nowadays for investigating the system's behavior in the industry and also in other fields of living. Especially nowadays, when the computation power of today's personal computers is very high and the prize is relatively low the usability of the simulation grows.

The modelling stage tries to describe the system either mathematically or practically (Luyben 1989), (Maria 1997). The mathematical description for example uses material, heat etc. balances (Ingham et al. 2000) depending on the type of the system, whether it is chemical reactor (Russell and Denn 1972), heat exchanger or electric motor. On the other hand, real model is usually small representation of the originally nonlinear system and we expect that results of experiments on this model are also valid or comparable to those on the real system. The big advantage of the modelling is in its safety – experiments on some real systems could be sometime hazardous. Nevertheless, experiments on the real or abstract model are usually much cheaper than those on the original system which is sometimes big and components are expensive.

This contribution combines two modelling techniques. At first, the mathematical model of the water tank will

be derived, then simulations were done on this model and results are verified by measurements on the real model of the water tank as a part of the Armfield's Process Control Teaching System PCT40. This real model represents the second modelling approach.

The mathematical model of the water tank system is mathematically described by the first order nonlinear Ordinary Differential Equation (ODE) (Luyben 1989). The simulation of this model consists of static and dynamic analyses.

The static analysis means solving of this ODE in the steady-state, i.e. the derivatives with the respect to time are equal to zero (Ingham et al. 2000). The nonlinear ODE is then reduced to the nonlinear algebraic equation which can be solved for example with the use of simple iteration methods (Saad 2003). The result of the static analysis could be optimal operating point or the range where the input variable could vary from the practical point of view.

On the other hand, the dynamic analysis observes the behavior of the system after the step change of the input quality, in this case the change of the feed volumetric flow rate inside the water tank. The dynamic analysis means mathematically the use of some numerical methods for solving of the ODE. The main groups of numerical methods are one-step methods for example Euler's method, Runge-Kutta's method, or multi-step methods Predictor-Corrector etc. (Johnston 1982). The advantage of these methods is that they are easily programmable even more they are build-in functions in the mathematical software like Matlab (Mathews and Fink 2004), Mathematica etc. (Kaw et al. 2014).

The contribution is divided into four main parts. The first part is introduction, next the modelling procedure is discussed from the theoretical point of view in the second part. The third part applies the procedure to the real model of the water tank and the last part is conclusion.

All simulations were done in Matlab, version 7.0.1.

SIMULATION PROCEDURE

As it is written above, this paper will describe the modelling and simulation procedure which usually precedes the design of the controller. This procedure could be generally divided into 6 parts which are displayed in Figure 1. Each part is important for the designing of the accurate model.

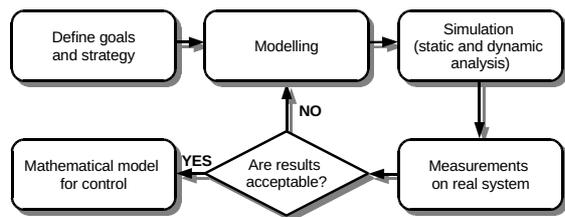


Figure 1: General modelling procedure

Goals and Strategy Definition

The first step is typically dedicated to the collection of all available information about the system. It defines the input, state and output variables and also constants and parameters of the system. Then, the output variable or variables which are important for control are chosen together with the most suitable input variables which could be used for the control.

In some cases, not every input variable can be used for control from the practical point of view. For example, the input concentration of the reactant in the chemical reactor is typical input variable but it is not very useable for control – it is hard change the input concentration quickly. The choice of the output variable is very similar – some output variables are not easily measurable.

This part of the procedure employs control engineers that have experience with the choice of the input and output variables together with process engineers which know the system from the process point of view.

Modelling

While all variables and relations between them are collected we can move on to the description of the system in some way – we collect a model of the observed plant.

There are two main types of models – *physical (real) models* and *abstract models*. The real model is represented by the copy of the system, usually small or similar to the original one. On the other hand, the mathematical model is usually used as an abstract model of the system.

The real system could be often nonlinear, unstable – generally very complex or partly misunderstood. The mathematical description of all quantities and relations between them lead to very complex and mostly insoluble mathematical model. Thus mathematical models do not strictly describe all the properties and relations inside the system, but pick up the most important ones and introduce constants and simplifications which reduce the complexity of the system.

Common simplifications could be found assumptions that volumes, heat capacities etc. are constant during the measurements. In some cases they are not constant but its changes are negligible. On the other hand, too many simplifications could lead to very simple mathematical

model behavior of which is different from the real system. To find compromise between the simpler but proper mathematical model are the most important part of modelling.

One tool which is employed here are balances inside the system. There are several types of balances – a material, a heat etc. The material balance in the steady-state, e.g. in state where state variables are steady and do not change, can be generally described in the word form in

Figure 2:



Figure 2: The word form of the mass balance in the steady-state

Unfortunately, most of the variables vary in time and steady-state balance is not suitable. We can introduce the dynamic material balance which contains changes with respect to time in the form of the accumulation – see the word equation in Figure 3:

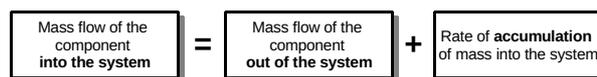


Figure 3: The word form of the mass balance in the dynamics

The collection of all balances inside the system results in one or more linear or nonlinear mathematical model usually in the form of algebraic or differential equations.

Simulation

Ones we have the mathematical model of the system we can observe the behavior of the system in the steady-state and the dynamics. It means that the mathematical model is solved with the use of iterations methods or numerical methods for solving of differential equations. The *steady-state analysis* for stable systems involves computing values of state variables in time $t \rightarrow \infty$, when changes of these variables are equal to the zero. That means that all equations which consist of derivations with the respect to zero have these derivations equal to the zero, i.e.

$$\frac{d(\bullet)}{dt} = 0 \quad (1)$$

There are many methods for solving of this problem. If the system is linear, the set of differential equations can be rewritten to the set of linear equations which can be solved by general, well known, methods like matrix-inversion, Gauss elimination etc. or with the use of some types of iterative methods. However, the most of processes are nonlinear which leads us to the set of nonlinear equations. Despite the fact that there is a

possibility of the analytical solution, iterative methods are used more often.

For example, the *simple iterative method* (Saad 2003) is often used for solving of nonlinear equations. This method leads to the exact solution for an appropriate choice of initial iteration and for the fulfilled convergence condition. Its advantage is that it does not need special modifications and side calculations according to other iterative methods like Newton's method etc. Although this method converges slower than Newton's method, this disadvantage is unimportant nowadays, when the speed of computers is very high. This method will be used for solving of a steady-state.

The second, *dynamic analysis* uses results from the steady-state as an initial conditions and solve mathematical model, usually in the form of one or more differential equation. Systems where state variable are dependent only to the one variable, for example time, are called lumped-parameters systems. Mathematical model of these systems is described by ordinary differential equations (ODE). The second types are systems, where state variable depends on more than one variable – e.g. time and space variable and the mathematical model consists of partial differential equations (PDE).

There are a lot of numerical methods for solving of differential equations, such as an Euler method, Runge-Kutta's methods, a predictor-corrector method etc. The advantage of these methods is that they have good theoretical background, modifications and even more they are mostly build-in functions in mathematical software such as Matlab (Mathews and Fink 2004) or Mathematica (Kaw et al. 2014).

Measurements on the Real System and Verification

Important part is the verification of the abstract mathematical model by reference measurements on the real system or its model. These experiments show accuracy of the mathematical model. The best way is to do the measurements for the same values and conditions on the real system and the mathematical model and then compare results if they are acceptable or we must recollect the mathematical model in the different way or take into the account some of assumptions made in the previous step.

This step is not feasible in every case but the mathematical model without this verification is not 100% trustworthy.

Mathematical Model for Control

If the mathematical model describes the system in proper way we can continue with the choice of input and output variables and the optimal control strategy. The simulation of the dynamic behavior could help us for example in the choice of the External Linear Model (ELM) in the adaptive control (Bobal et al. 2005), (Vojtesek and Dostal 2012).

REAL MODEL – WATER TANK

The procedure described in the previous chapter was tested on the real model of the water tank which is one part of the Multifunctional process control teaching system PCT40 from Armfield – see Figure 4. This equipment includes also other models of processes such as Continuous Stirred Tank Reactor (CSTR) or heat exchanger.

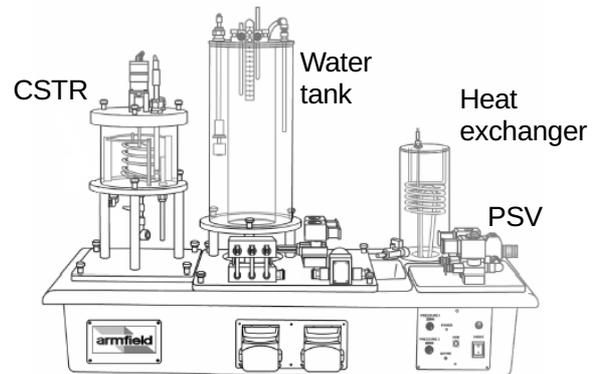


Figure 4: Multifunctional process control teaching system PCT40

Goals and Strategy Definition

This system combines both modelling techniques – it is small representation of the water tank with the volume of 4-liter original of which is usually much bigger with huge volume. The mathematical model of this system could be also easily derived. The schematic representation of the water tank can be found in Figure 5.

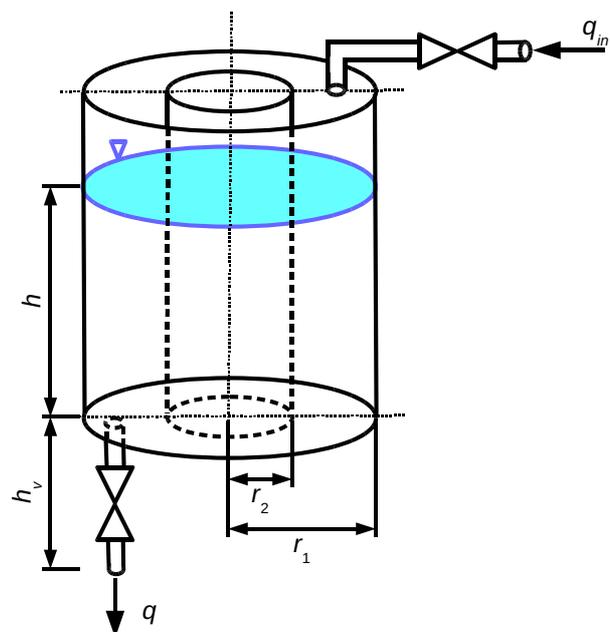


Figure 5: Schematic representation of the water tank

The model consists of plastic transparent cylinder with inner radius $r_1 = 0.087 \text{ m}$. There is another plastic

transparent cylinder inside due to quicker dynamic response of the system lower usage of feeding water. The outer radius of this smaller cylinder is $r_2 = 0.057 \text{ m}$ and the maximal water level in the tank is $h_{max} = 0.3 \text{ m}$.

In the Figure 5, q denotes the volumetric flow rate, h is used for the water level and r are radiuses of inner and outer cylinders. The input variable is the volumetric flow rate of the feeding water q_{in} and state variables are water level h in the tank and output volumetric flow rate of the water which comes from the tank, q .

The goal of the modelling is to create the mathematical model which describes dependence of the water level, h , on the input volumetric flow rate, q_{in} .

Modelling of the Water Tank

The modelling uses material balance described in the general word form as in Figure 3. In this concrete case it could be rewritten to the word equation displayed in Figure 6.

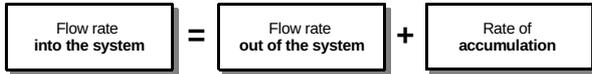


Figure 6: Material balance inside the water tank

Which is mathematically:

$$q_{in} = q + \frac{dV}{dt} \quad (2)$$

where V is a volume of the water inside the tank and t is used for the time.

The volume of the tank is generally

$$V = F \cdot h \quad (3)$$

for F as a area of the base due to cylindrical shape of the tank. It means, that balance (2) could be rewritten to the form

$$q_{in} = q + F \cdot \frac{dh}{dt} \quad (4)$$

where F is in this case

$$F = \pi \cdot r_1^2 - \pi \cdot r_2^2 = 1.36 \cdot 10^{-2} \text{ m}^2 \quad (5)$$

It is also known, that volumetric flow rate through the water valve is nonlinear function of the water level, i.e.

$$q = k \cdot \sqrt{h} \quad (6)$$

where k is a valve constant which is specific for each valve and depend on the geometry and type of the valve.

If we put equation (6) inside (4) the resulting mathematical model is:

$$\frac{dh}{dt} = \frac{q_{in} - k \cdot \sqrt{h}}{F} \quad (7)$$

There should be introduced one simplification – the height of the discharging valve, h_v in Figure 5, is neglected.

The unknown constant k could be computed for example from the steady state (variables with superscript (\cdot^s)), where $q_{in}^s = q^s$ and equation (6) is

$$q^s = k \cdot \sqrt{h^s} \Rightarrow k = \frac{q^s}{\sqrt{h^s}} \quad (8)$$

The water tank is fed via Proportioning Solenoid Valve (PSV) which could be operated in the range 0 – 100%. This range is practically 0 – $2.5 \cdot 10^{-5} \text{ m}^3 \cdot \text{s}^{-1}$.

We have made measurements on the real model for the 60% of valve operation which represents input flow rate $q_{in} = 1.5 \cdot 10^{-5} \text{ m}^3 \cdot \text{s}^{-1}$. The result of the measurement is shown in Figure 7.

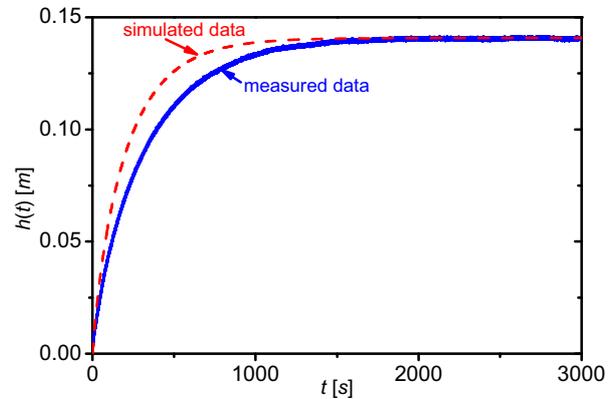


Figure 7: Measured and simulated data for $k = 4.01 \cdot 10^{-5}$ and $q_{in} = 1.5 \cdot 10^{-5} \text{ m}^3 \cdot \text{s}^{-1}$

The final (steady-state) value of the water level h is for this flow rate $h^s = 0.141 \text{ m}$. It means, that the valve constant k is

$$k = \frac{q_{in}}{\sqrt{h^s}} = \frac{1.5 \cdot 10^{-5}}{\sqrt{0.141}} = 4.0107 \cdot 10^{-5} \quad (9)$$

The mathematical model (7) is now complete and we can move on to simulation analyses.

Simulation and Verification of the Model

The simulation is very often connected to the verification part because it is good to know if the derived mathematical model is accurate enough.

The result of the first simulation analysis for the same input volumetric flow rate $q_{in} = 1.5 \cdot 10^{-5} \text{ m}^3 \cdot \text{s}^{-1}$ is shown in Figure 7 – the dashed line. It is clear, that although simulated and measured outputs reaches the same final value, the dynamics is much different – the mathematical model has quicker output response.

There were done five more reference measurements for different volumetric flow rate and the results for the final value and values of the valve constant, k , are shown in Table 1.

Table 1: Results of reference measurements on the real model

Flow rate $q_{in} [m^3 \cdot s^{-1}]$	Steady-state water level $h^s [m]$	Valve constant $k [-]$	New valve constant $k_n [-]$
$1.34 \cdot 10^{-5}$	0.095	$4.346 \cdot 10^{-5}$	$3.239 \cdot 10^{-5}$
$1.43 \cdot 10^{-5}$	0.118	$4.153 \cdot 10^{-5}$	$3.239 \cdot 10^{-5}$
$1.50 \cdot 10^{-5}$	0.141	$4.011 \cdot 10^{-5}$	$3.233 \cdot 10^{-5}$
$1.68 \cdot 10^{-5}$	0.195	$3.804 \cdot 10^{-5}$	$3.227 \cdot 10^{-5}$
$1.86 \cdot 10^{-5}$	0.258	$3.658 \cdot 10^{-5}$	$3.216 \cdot 10^{-5}$
$1.93 \cdot 10^{-5}$	0.285	$3.618 \cdot 10^{-5}$	$3.215 \cdot 10^{-5}$

It can be seen that resulted values of the valve constant, k , in Table 1 vary in relatively big range $3.618 \cdot 10^{-5} - 4.346 \cdot 10^{-5}$ which produces very inaccurate results – similar as in Figure 7. The reason for these inaccurate and very different values can be found in the simplification introduced in the modelling part, where the height of the discharging valve was neglected. This height $h_v = 0.076 m$ has, of course, impact to the dynamics of the system and also to the valve constant. Table 1 also shows in the last column recomputed values of valve constant, k_n , for the measurements, where the height of the valve is taken into the account. These values are very close to each other and output responses of the mathematical model with this new constant k_n are much closer to the measured ones – see Figure 8 which presents results for $1.5 \cdot 10^{-5} m^3 \cdot s^{-1}$ (i.e. 60% of maximal q_{in}) and $1.93 \cdot 10^{-5} m^3 \cdot s^{-1}$ (i.e. 78%).

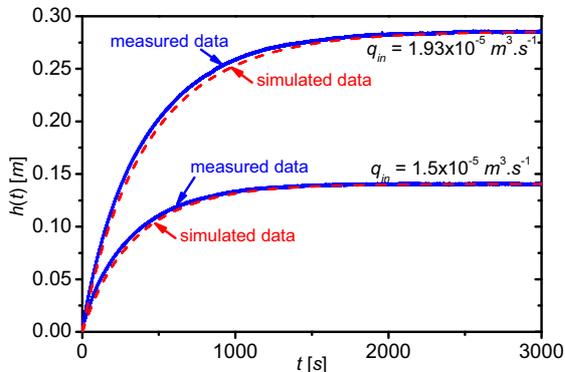


Figure 8: Measured and simulated data for $k_n = 3.23 \cdot 10^{-5}$ and $q_{in} = 1.5 \cdot 10^{-5} m^3 \cdot s^{-1}$

As a result, the mean value of the new valve constant, k_n , is taken into account for the next computations, i.e.

$$k_n = 3.2282 \cdot 10^{-5} \quad (10)$$

The height of the valve h_v is then reflected in the new valve constant, but the water level in the next analyses is measured from the bottom of the water tank because all measuring devices have an zero water level at the floor of the tank – a proportional pressure sensor for accurate measuring and the reference visual scale at the cover of the plastic tank.

The Steady-state Analysis.

The steady-state analysis means that we solve the mathematical model with the condition $d(\cdot)/dt = 0$, i.e. ODE (7) is transferred to the nonlinear algebraic equation:

$$h^s(q_{in}) = \left(\frac{q_{in}}{k} \right)^2 \quad (11)$$

where the optional variable is the input volumetric flow rate, q_{in} . There were done simulation analysis for the range $q_{in} = \langle 0; 2.5 \cdot 10^{-5} \rangle m^3 \cdot s^{-1}$ and results are shown in the Figure 9.

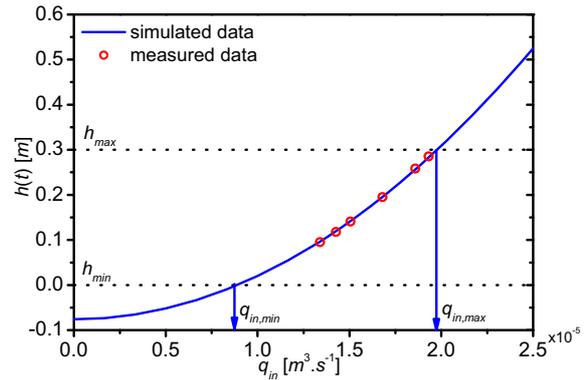


Figure 9: The steady-state analysis of the mathematical model

This analysis shows nonlinear behavior of the system and also we can choose the volumetric flow rate in the range $q_{in} = \langle 8.86 \cdot 10^{-6}; 1.98 \cdot 10^{-5} \rangle m^3 \cdot s^{-1}$ because lower value of q_{in} means that we did not get enough water in the tank and vice versa – the flow rate bigger than $q_{in} = 1.98 \cdot 10^{-5} m^3 \cdot s^{-1}$ results in bigger water level than its maximal value h_{max} . Red dots in the Figure 9 display results of measured steady-state value of the water level from Table 1.

The Dynamic Analysis.

The dynamic analysis solves the ODE with the use of some numerical methods. In this case, the Runge-Kutta's standard method was used because it is easily programmable and even more it is build-in function in used mathematical software Matlab. The working point was characterized by the input volumetric flow rate $q_{in}^s = 1.5 \cdot 10^{-5} m^3 \cdot s^{-1}$ which is somewhere in the middle of the operating interval defined after the static analysis in the Figure 9.

The input variable, $u(t)$, is the change of the initial q_{in}^s and the output variable is the water level in the tank. The input and the output variables are then generally:

$$u(t) = \frac{q_{in}(t) - q_{in}^s}{q_{in}^s} \cdot 100 [\%]; y(t) = h(t) [m] \quad (12)$$

The simulation time was 3000 s, six step changes of the input variable $u(t)$ were done and results are shown in Figure 10.

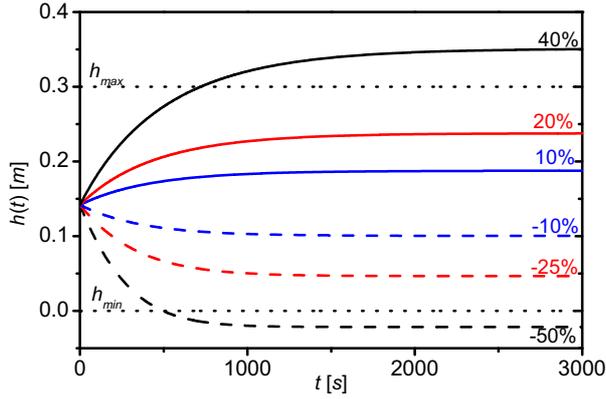


Figure 10: The dynamic analysis for various step changes of the input volumetric flow rate q_{in}

Output responses show that this output has asymmetric responses – the final value is different in sign and also in order for positive and negative step changes. Even more, for it is inappropriate to choose the input step change of the volumetric flow rate lower than approximately -41% and bigger than +31% because the resulted water level is lower or higher than physical properties of the water tank.

Mathematical Model for Control

The last step in the procedure defined in the theoretical part is description of the system from the control point of view. This description depends on the chosen control strategy. For example, one strategy of an adaptive control uses the External Linear Model (ELM) of the originally nonlinear system for construction of the adaptive controller parameters of which are recomputed in each sampling period according to the recursively identified parameters of the ELM (Vojtesek and Dostal 2012).

In this case, all output responses in Figure 10 could be expressed by the first or the second order transfer functions (TF), for example in the continuous-time

$$G_1(s) = \frac{b(s)}{a(s)} = \frac{b_0}{s + a_0} \quad (13)$$

$$G_2(s) = \frac{b(s)}{a(s)} = \frac{b_1s + b_0}{s^2 + a_1s + a_0}$$

or in the discrete time

$$G_1(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_1z^{-1}}{1 + a_1z^{-1}} \quad (14)$$

$$G_2(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_1z^{-1} + b_0z^{-2}}{1 + a_1z^{-1} + a_0z^{-2}}$$

We can do now simple least-squares method for the off-line identification of the simulated data from the dynamic analysis to investigate parameters of polynomials $A(z^{-1})$ and $B(z^{-1})$ from the (14). The qualitative criterion S_e in this case is sum of squared

differences between the simulated output y_{sim} and the identified output y_{id} :

$$S_e = \sum_{i=1}^N (y_{sim}(i) - y_{id}(i))^2 \quad [m^2] \quad (15)$$

where N is a number of steps, i.e. $N = T_f/T_v$ when T_f is final time and T_v is sampling period. Results of this off-line identification for both the first and the second order transfer functions for example for step changes $u(t) = -40$ and $+50$ % are shown in Figure 11 and Figure 12.

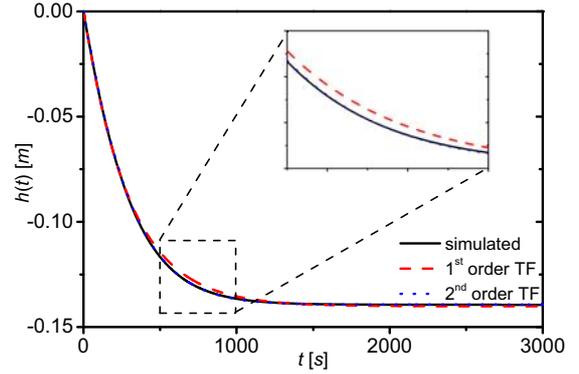


Figure 11: Results of off-line identification for step change $u(t) = -40\%$

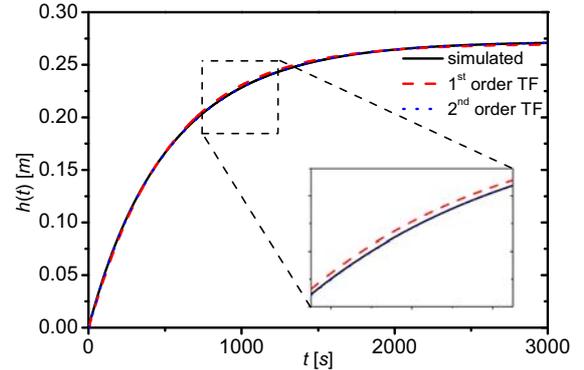


Figure 12: Results of off-line identification for step change $u(t) = +50\%$

It is clear that both TF in (14) describes simulated data relatively well, visually worse course is for the 1st order TF for negative step change $u(t) = -40\%$ - see detailed cuts in Figure 11 and Figure 12.

Table 2 shows values of S_e for both 1st and 2nd order TF for all step changes. We can say, that the 2nd order TF describes the controlled output in more accurate way.

Table 2: Computed quality criterion S_e for 1st order and 2nd order transfer function (TF)

$u(t)$ [%]	1 st order TF S_e [m^2]	2 nd order TF S_e [m^2]
-40	$3393.30 \cdot 10^{-6}$	$16.28 \cdot 10^{-6}$
-20	$237.46 \cdot 10^{-6}$	$36.35 \cdot 10^{-6}$
-10	$21.71 \cdot 10^{-6}$	$9.27 \cdot 10^{-6}$
+10	$21.71 \cdot 10^{-6}$	$9.95 \cdot 10^{-6}$
+25	$500.54 \cdot 10^{-6}$	$66.01 \cdot 10^{-6}$
+50	$6555.60 \cdot 10^{-6}$	$308.80 \cdot 10^{-6}$

CONCLUSIONS

The goal of this contribution was to show the procedure of modelling and simulation before the design of the controller. The system properties together with the most important quantities and relations between them are sketch out, then the mathematical model was derived with the use of balances inside the system and finally the steady-state and dynamic analyses were done to obtain the behavior of the system.

Important part of the modelling is the verification of the simulated data on the real system or its model. This comparison shows an accuracy of the mathematical model. The procedure was tested on the real model of the water tank as a part of laboratory equipment. The first simulation studies have shown that introduced simplification leads to inaccurate results. The height of the discharging valve, which was previously neglected, has affected the value of the valve constant and consequently the course of the output in the significant way and it must be taken into the account. The steady-state analysis produces the range of the input volumetric flow rate in which the measurements have practical meaning. The second, dynamic, analysis has shown that the output could be described rather by the second order transfer function than the first order one because of the accuracy of the description. The next work will be focused on the choice of the optimal control strategy, simulation experiments and again verifications on the real model.

REFERENCES

- Armfield: Instruction manual PCT40, Issue 4, February 2005
- Bobal, V.; J. Böhm; J. Fessler; J. Machacek. 2005 *Digital Self-tuning Controllers: Algorithms, Implementation and Applications*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag London Limited. 2005, ISBN 1-85233-980-2.
- Ingham, J.; I. J. Dunn; E. Heinzle; J. E. Prenosil. 2000 *Chemical Engineering Dynamics. An Introduction to Modeling and Computer Simulation*. Second. Completely Revised Edition. VCH Verlagsgesellschaft. Weinheim, 2000. ISBN 3-527-29776-6.
- Johnston, R. L. 1982. *Numerical Methods*. John Wiley & Sons. ISBN 04-718-666-44
- Kaw, K.; C. Nguyen; L. Snyder. 2014. *Holistic Numerical Methods*. Available online <<http://mathforcollege.com/nm/>>
- Luyben, W. L. 1989. *Process Modelling, Simulation and Control for Chemical Engineers*. McGraw-Hill, New York 1989. ISBN 0-07-039-1599
- Maria A. 1997. Introduction to modeling and simulation. In: *Proceedings of the 1997 Winter Simulation Conference*. 7-13.
- Mathews, J. H.; K. K. Fink. 2004. *Numerical Methods Using Matlab*. Prentice-Hall. ISBN 01-306-524-82.
- Russell, T.; M. M. Denn. 1972. "Introduction to chemical engineering analysis". New York: Wiley, 1972, xviii, 502 p. ISBN 04-717-4545-6.
- Saad, Y. 2003. *Iterative Methods for Sparse Linear Systems*. Society for Industrial & Applied. ISBN 08-987-153-42.
- Vojtesek, J.; P. Dostal. 2012 Simulation of Adaptive LQ Control of Nonlinear Process. *Studies in Informatics and Control*, Vol. 21, Issue 3, Bucharest, Romania, 2012, p. 315-324. ISSN 1220-1766

AUTHOR BIOGRAPHIES



JIRI VOJTESEK was born in Zlin. Czech Republic and studied at the Tomas Bata University in Zlin. where he got his master degree in chemical and process engineering in 2002. He has finished his Ph.D. focused on Modern control methods

for chemical reactors in 2007. His email contact is vojtesek@fai.utb.cz.



PETR DOSTAL studied at the Technical University of Pardubice. He obtained his Ph.D. degree in Technical Cybernetics in 1979 and he became professor in Process Control in 2000. His research interest are

modeling and simulation of continuous-time chemical processes. polynomial methods. optimal. adaptive and robust control. You can contact him on email address dostalp@fai.utb.cz.



MARTIN MASLAN was born in Uherske Hradiste. He finished Tomas Bata University in Zlin with a Bachelor degree in 2012 and he is studying last year of Master degree in Automatic Control and Informatics at this university. His interest

in this field include measurement and control and he would like to deal with it in the future. His email contact is martin.maslan@seznam.cz.

THE EFFECT OF INITIAL ESTIMATED POINTS ON OBJECTIVE FUNCTIONS FOR OPTIMIZATION

Mahdi Soltani
Annalisa Pola

Dip. di Ingegneria Meccanica e Industriale (DIMI)
Università degli Studi di Brescia
Brescia, Italy
mahdi.soltani@unibs.it
annalisa.pola@unibs.it

Qiang Xu

School of Computing and Engineering
The University of Huddersfield
Huddersfield, UK
Q.Xu2@hud.ac.uk

KEYWORDS

Component, Optimization, initial estimated points, Creep constitutive Equations.

ABSTRACT

Research progress on the optimization in order to obtain the value of material constants for a set of creep damage constitutive equations was presented, including (1) a brief review of continuum creep damage modeling and the designs of objective functions; (2) case study reporting the influence of the initial start points on the final results; and (3) discussion and conclusion. As far as the authors know, there is no any published paper addressing specifically on the influence of starting values on the final results. In order to overcome the difficulty or inaccuracy it is suggested in this paper to (1) check the accuracy of a particular set of experimental data, (2) review the method to depict the relationship between the stress level and minimum strain rate, (3) to design a better objective functions so that the convergence is ensured to all the stress levels.

INTRODUCTION

Constitutive equations, the mathematical description of material behaviour, are crucial for academic and industry. Hence, it is extremely important to be able to determine the best values for the material constants for a given set of constitutive equations

The aim of optimization processes is to determine the material constants of a set of constitutive equations with the best fits experimental data, typically using the least squares method. The principle of the least squares approximation to optimize a problem minimizes the sum of the squares of deviation between predicted values and experimental values (Ortega, 1994) (whittle, 1984).It could supply an improved procedure to determine "Direction Vector" to generate a sequence of points by the use of optimization-technique introduced by Numerical algorithms group (NAG) Library.

Creep damage mechanics has been developed and applied for dealing with the creep damage failure in high temperature industries where the core is the development of a set of creep damage constitutive equations. It is in this context, the determination of creep damage constants has been an issue for research community.

The creep experimental data described by (Hayhurst, et al., 2003) for 316 stainless steel at 550°C for six stress levels of 320, 300, 280 MPa (as high stress levels) and 231.65, 200.78, 169.6 MPa (as low stress levels) are used in the optimization and analysis with special initial estimated points (Table 1).

Table 1: Constant Starting Points for Objective functions (OF) within optimization process for 316 stainless steel at 550°C

	<i>A</i>	<i>B</i>	<i>H*</i>	<i>h</i>
<i>IEP</i>	1.53758E-07	0.009979478	0.50004175	47743.527

Although a recent research addressed that objective functions algorithms for optimization could be commenced from any unsystematic starting values without manual selection of initial points (Lia, et al., 2002), this paper tries to illustrate how change in initial Estimated Points (IEP) could lead to changes in the calculated constant material within creep constitutive equations (Table 2). Change in IEP in objective functions (OF) for components under lower stress levels, primarily, have experienced far less marked rather than the higher stress levels, as changing IEP (even less than 10%) could have a significant effect on the creep deformations under high stress levels. Table 2 shows 4 calculated models in different situation:

1. IEP increase by 10%; (*A*, *B*, *H**, *h*)
2. IEP increase by 10%; just only in *A* and *B*
3. IEP decrease by 10%; (*A*, *B*, *H**, *h*)
4. IEP decrease by 10%; just only in *A* and *B*

Table 2: Change in Starting Points for Objective functions (OF) within optimization process

<i>Model</i>	<i>A</i>	<i>B</i>	<i>H*</i>	<i>h</i>
base	1.53E-07	0.009	0.5	4.77E+04
1	1.69E-07	0.010	0.55	5.25E+04
2	1.69E-07	0.010	0.5	4.77E+04
3	1.38E-07	0.008	0.4	4.30E+04
4	1.38E-07	0.008	0.5	4.77E+04

CONTINUUM DAMAGE MECHANISMS BASED MODELLING

The material science and engineering research have been established a reasonable understanding of the major creep deformation and damage mechanisms. Nevertheless, according to each class of materials, there are various approaches to use the different mechanisms for creep damage constitutive and evaluation equations have been formulated of the power law functions of stress to give an account of the creep behaviour under uniaxial and multi-axial states of stress over the various temperature ranges. There are several uniaxial creep design methods basing on the Continuum Damage Mechanisms (CDM). CDM-based methods have been used to predict creep damage, the lifetime of creep deformation, the residual strength and rupture strain by utilizing the various constitutive equations.

Mechanisms-Based Model

Mechanisms-Based Model has a capability to utilize, practically, an alternative solution of the prominent problems in creep behaviours. Hence, the continuum approaches has merit of implementing a numerical method to simulate the processes of damage evolutions during the development in finite elements methods. (Barboza, et al., 2004) One of the proposed models is based upon the relation between Sine hyperbolic stress law and the impression velocities is pertinent to the technical analysis of creep deformations in the different ranges of stress and temperature (Yang, et al., 1995). The mechanisms-based model identified by the reproduction kinetics of dislocation motion using the finite element methods introduced for different materials such as for 0.5Cr-0.5Mo-0.25V ferritic steel under uniaxial over the temperature range of 600-675°C (Perrin & Huyhurst, 1996).

The effect of strain hardening through the primary-creep, H , has been formulated on the constitutive equations. Moreover, the cavities nucleate often occur on the grain boundaries so there is another variable introduced damage state, ω , that it is the inter-granular creep constrained cavitation damage. Furthermore, the effects of temperature on the creep constitutive equation have been considered (equation 1).

$$\begin{cases} \dot{\epsilon} = \frac{A}{(1-\omega)^n} \sinh[B\sigma(1-H)] \\ \dot{H} = \frac{h\dot{\epsilon}}{\sigma} \left(1 - \frac{H}{H^*}\right) \\ \dot{\omega} = D\dot{\epsilon} \\ n = B\sigma(1-H) \coth(B\sigma(1-H)) \end{cases} \quad (1)$$

Where A , B , H^* and h are introduced as the creep constant material. The numerical optimization methods need to be used to determine the material constants with the best fits experimental data for finding the best optimum of a problem by using the least squares method. In many problems, it is trying to approximate a set of data by a best fit (Perrin & Hayhurst, 1996) (Perrin & Huyhurst, 1996). Constitutive Equations corresponding

to the creep behaviours under multi-axial stress could be completed.

Kachanov-Rabotnov Model

The theory of creep damage mechanics was developed by Kachanov and Rabotnov in the 1960s that it has applied for analysing of creep rupture in different range of materials, although it has been modified within these years. They established a modern procedure introducing new parameters which describe the “continuity” (Ψ) and the “damage” (ω) of the components (MacLachlan & Knowles, 2001). There are several different modified finite elements based on K-R creep model in the various mechanical software packages such as ANSYS and ABAQUS (Ling, et al., 2007).

OPTIMIZATION MODE

The least squares optimization concentrated more on the stress levels (high stress levels and low stress levels) have been proposed on the suggested different creep constitutive equations. The objective function is usually specified to optimize a problem minimizing the sum of the squares of the deviation between predicted and experimental values. The following will demonstrate the put forward points more.

Objective function

Reference (Lia, et al., 2002) characterized the difficulties in determining the material constants via the principle of the least squares approximation experimental data. Moreover, it was expected that the Objective functions algorithms for optimization could be commenced from any unsystematic starting values inasmuch manual selection of initial points do not need to be used, hence, the objective functions carry out better than binary genetic algorithms suggested previously (Lin & Yangb, 1999). It is introduced two ways for determining the objective functions for optimization:

Objective function I (OFI):

The errors could be determined by the shortest distance between calculated and experimental data. Meanwhile, in order to develop the sensitivity, one extra term needs to be introduced (Lia, et al., 2002). Thus the objective function, $f(X)$, is defined by:

$$f(X) = \sum_{j=1}^{n_1} \sum_{i=1}^{m_j} w_{ij} d_{ij}^2 + \sum_{j=1}^{n_1} W_j \left(t_{m_j}^t - t_{m_j}^m \right)^2 \quad (2)$$

We need to introduce the definitions of the different terms used in the above equation;

Objective function II (OFII):

Where the errors could be determined by time only:

$$F(X) = \sum_{j=1}^{n_1} \sum_{i=1}^{m_j} W_j \left(t_{m_j}^t - t_{m_j}^m \right)^2 \quad (3)$$

The Overall Optimization for Creep Constitutive Equation

The recommended Objective functions includes the integration of square of the residual value of predicted and experimental strain for each points corresponding to different stress levels (equation 4), and for increasing the sensitivity, two extra terms are introduced, the residual value of computational and experimental lifetimes and minimum creep-strain rates (equation 5) [9]. The Overall Optimization for Creep constitutive Equation identified by (Hayhurst, et al., 2003):

- The Least Square optimization:

$$LS = \sum_{i=1}^{MS} \left\{ \left[\sum_{j=1}^{a_i} (\varepsilon_j^{\text{pred}} - \varepsilon_j^{\text{exp}})^2 \right]_i + Z_i (t_i^{\text{pred}} - t_i^{\text{exp}}) / t_i^{\text{exp}} + \alpha_i \{ (\dot{\varepsilon}_{\text{min}}^{\text{pred}} - \dot{\varepsilon}_{\text{min}}^{\text{exp}}) / \dot{\varepsilon}_{\text{min}}^{\text{exp}} \}_i \right\} \quad (4)$$

- The modified Least Square optimization:

$$LS^* = \sum_{i=1}^{MS} \left\{ \left[\sum_{j=1}^{a_i} \left(\frac{\varepsilon_j^{\text{pred}} - \varepsilon_j^{\text{exp}}}{\varepsilon_{a_i}^{\text{exp}}} \right)^2 \right]_i + Z_i (t_i^{\text{pred}} - t_i^{\text{exp}}) / t_i^{\text{exp}} + \beta_i \{ (\varepsilon_f^{\text{pred}} - \varepsilon_f^{\text{exp}}) / \varepsilon_f^{\text{exp}} \}_i + \alpha_i \{ (\dot{\varepsilon}_{\text{min}}^{\text{pred}} - \dot{\varepsilon}_{\text{min}}^{\text{exp}}) / \dot{\varepsilon}_{\text{min}}^{\text{exp}} \}_i \right\} \quad (5)$$

If all parts of the creep deformation curves for each stress levels (High stress and also low stress with long lifetime) have been available to investigate, therefore, the least-square optimization will be able to improve greatly; in this situation, different stages of the creep curves could be conventionalized with respect to the experimental-values at the end of each of the corresponding stages, thus the new suggested optimization model is divided into different parts which each part will be pertinent to one stage of the creep curves. (Primary and secondary creep curves) (2005).

Changing Accuracy in Optimization Process

Any changes in accuracy with different starting points lead to marked changes in calculated constant materials. This study try to illustrates the information about percentage of changes in computed results with different accuracy demanded after 10% increase or decrease in starting points either A and B or all four creep constant materials (A, B, H* and h).

In this situation, there are some different ‘‘Amplification factors’’ which could control optimization process. Amplification factors could supply an improved procedure to determine ‘‘Direction Vector’’ to generate a

sequence of points by the use of optimization - techniques implanted in the NAG routine.

Albeit, the accuracy of linear minimizations leads to decrease the number of iterations executed by NAG, it would cause an increase in the number of invoked subroutine relevant to Least Square optimization (NAG, 2011). In this situation, the number of iterations required depends on:

- the number of variables of constitutive equations,
- the number of the residual value of computational and experimental data
- the behaviour of $F(X)$ that $X = x^{(k)} + \alpha^{(k)} p^{(k)}$,
- the accuracy required,
- the distance of the starting point from the solution.

NAG library has determined two different parameters (ETA, XTOL) are the key factors to determine the accuracy demanded. It supposed to change the amount of two mentioned parameters, separately, to achieve a same point.

- ETA indicates how precisely the linear minimizations are to be performed. (Equation 4) is an approximate minimum with respect to $\alpha^{(k)}$; in this Subroutine, $\alpha^{(k)}$ plays a role in specifying how accurate in the linear-minimization is to be done.

$$F(x^{(k)} + \alpha^{(k)} p^{(k)}) \quad (6)$$

- XTOL indicates the preciseness in x to which the solution is required (NAG, 2011). If x_{true} is the true value of x at the minimum, then x_{sol} , the estimated position prior to a normal exit, is such that

$$\|x_{\text{sol}} - x_{\text{true}}\| < XTOL \times (1.0 + \|x_{\text{true}}\|) \quad (7)$$

NAG library introduces a controller parameter (is called ETA) that the user can specify the accuracy of minimum with respect to $\alpha^{(k)}$. Obviously, it will be located more precisely for small values of ETA (say 0.01) than for large values. Although it reduces the number of iterations carried out by NAG, it will increase the number of calls of Least-Square Optimization made. Changing in accuracy in liner minimizations (ETA) with different starting point at such as 0.5, 0.1, 0.01 and 1.0E-4 at a demanded accuracy in x such as 1.0E-4 lead to a noticeable change in calculated material constant (Table 3 and 4).

Table 3: The Percentage of Change between Initial Calculated Constant materials and calculated results with different IEP when ETA: 0.5(A), and ETA: 0.1(B), XTOL=1.0E-04

ETA=0.5 and XTOL=1.0E-04				
Model	A	B	H*	h
1	-22%	12%	-3%	-15%
2	-76%	17%	-23%	4%
Base	(A)			
3	20%	-3%	-7%	-1%
4	19%	-3%	-2%	-2%
ETA=0.1 and XTOL=1.0E-04				
Model	A	B	H*	h
1	-4%	5%	-1%	-32%
2	-22%	10%	-14%	-37%
Base	(B)			
3	18%	-2%	-7%	6%
4	31%	-6%	3%	-6%

Table 4: The Percentage of Change between Initial Calculated Constant materials and calculated results with different IEP when ETA: 0.01(c), and ETA: 1.0E-4(d); XTOL=1.0E-04

ETA=0.01 and XTOL=1.0E-04				
Model	A	B	H*	h
1	20%	-6%	13%	-21%
2	-8%	7%	-12%	-43%
Base	(C)			
3	31%	-8%	0%	-10%
4	26%	-7%	-2%	-26%
ETA=1.0E-4 and XTOL=1.0E-04				
Model	A	B	H*	h
1	23%	-11%	17%	-29%
2	-36%	14%	-18%	-27%
Base	(D)			
3	27%	-9%	1%	-22%
4	22%	-7%	-2%	-34%

ANALYSIS

Choosing more appropriate starting values for each material constant within the creep constitutive equations (equation 1) manage to provide a better possibility to converge to one of the local minimum than to a global one (Mustata & Hayhurst, 2005).

When some new experimental creep data (with two new stress levels; 240MPa, 185.3MPa) are added to the last mentioned collection, obviously, all of the previous calculated results will be changed. This paper tries to show that if there is a change in the previous starting points (even by 10 per cent) within this new collection, it could lead to changing significantly in the calculated constant material, especially in high stress levels (Figure 1).

The calculated material constants for creep constitutive equations under low stress levels (for example 231.65MPa, 200.78MPa, 185.3MPa and 169.67MPa), primarily, when initial estimated points for all constant material (A, B, H*, h) are changed by 10%, the final calculated strain will increase generally less than 15% thus it is expected that the optimization process could be started from any random starting points as the final result

especially in low stress level (Figure 2).

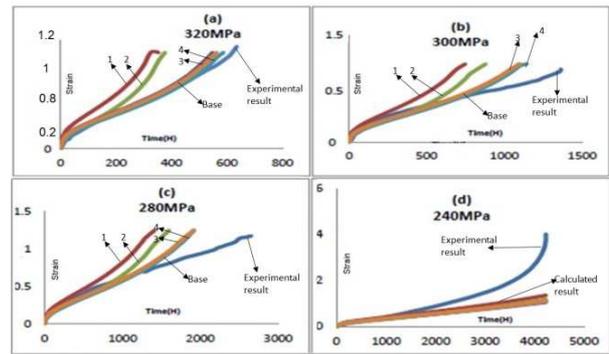


Figure 1: Creep Deformation Curves with Different IEP under higher Engineering Stress Levels

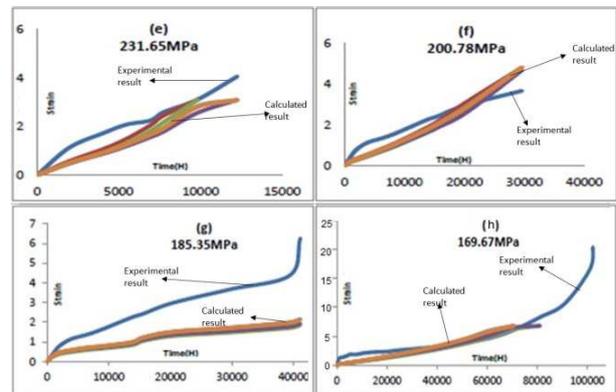


Figure 2: Creep Deformation Curves with Different IEP under lower Engineering Stress Levels

However, this effect is much considerable, especially, for high stress levels (for example 320MPa, 300MPa, 280MPa and 240MPa). The final calculated strain will increase primarily more than 25% (such as 320MPa, at the tertiary creep stages, changes accounted for a significant increase by even more than 50%).

As the creep primary stage for high stress levels is more significant than that under lower stress levels, the reduction of percentages in calculated variables corresponding to the strain hardening through the primary-creep for higher stress levels has a consequential effect on the calculated strain and creep deformation curves (Figure 1). Therefore, these changes could not be neglected, thus using of different starting points lead to different results for objective functions within optimization process.

More fundamentally, careful observation of the numerical predicted creep curves shown in following Figures reveals that: the numerical predictions for high stress level are essentially larger than the experimental ones while the predicted creep curves for low stress level is essential underestimated. As long as the required accuracy for optimisation is achieved, the programme will stop resulting significantly different sets of final material constants. This consequently raises the question of how to design a better objective function to ensure the

agreement at different stress level is ensured rather than just the global agreement.

CONCLUSION

In this paper, the analysis of changing initial estimated points in optimization process has been proposed for different stress levels. This is achieved by using objective function within an advanced optimization techniques for a creep constitutive equations, which are as follows.

- Any changes in entered experimental data with the change in IEP could lead to a significant change in the calculated constants for materials within creep constitutive equations especially for material under higher stress levels.
- The effect of different accuracy in for the linear minimizations has been performed to approach to accurate local minima than a global minimum.

REFERENCES

- Barboza, M. J., Neto, C. M., and Silva, C. R. 2004. "Creep mechanisms and physical modeling for Ti-6Al-4V. *Materials Science and Engineering*", 369(1-2), 201-209.
- Hayhurst, D. R., Vakili-Tahami, F., and Zhou, J. Q. 2003. "Constitutive equations for time independent plasticity and creep of 316 stainless steel at 550 °C. *International Journal of Pressure Vessels and Piping*", 80(2), 97-109.
- Lia, B., Lin, J., and Yaoc, X. 2002. "A novel evolutionary algorithm for determining unified creep damage constitutive equations. *International Journal of Mechanical Sciences*", 44(5), 987-1002.
- Lin, J., and Yang, J. 1999. "GA-based multiple objective optimisation for determining viscoplastic constitutive equations for superplastic alloys". *International Journal of Plasticity*, 15(11), 1181-1196.
- Ling, X., Zheng, Y., You, Y., and Chen, Y. 2007. "Creep damage in small punch creep specimens of Type 304 stainless steel." *International Journal of Pressure Vessels and Piping*, 84(5), 304-309.
- MacLachlan, D. W., and Knowles, D. M. 2001. "Modeling and prediction of the stress rupture behaviour of single crystal super alloys." *Materials Science and Engineering*, 302(2), 275-285.
- Mustata, R., and Hayhurst, D. 2005. "Creep constitutive equations for a 0.5Cr 0.5 Mo 0.25V ferritic steel in the temperature range 565 °C–675 °C." *International Journal of Pressure Vessels and Piping*, 82(5), 363-372.
- NAG. 2011. "Numerical Algorithms Group". Tratto il giorno 03 15, 2011 da <http://www.nag.co.uk/index.asp>
- Ortega, J. 1994. "An introduction to FORTRAN 90 for science computing (1st ed.)." *Sunders College Publishing*.
- Perrin, I. J., and Hayhurst, D. R. 1996. "A method for the transformation of creep constitutive equations". *International Journal of Pressure Vessels and Piping*, 68(3), 299-309.
- Perrin, I. J., and Huyhurst, D. 1996. "Creep constitutive equations for a 0.5Cr 0.5 Mo 0.25V ferritic steel in the temperature range 565 °C–675 °C." *Journal of Strain Analysis*, 31(4).
- Whittle, P. 1984. "Prediction and Regulation by Linear Least-Square Methods (2nd ed.)." oxford: basil blackwell.
- Yang, F., Li, J. C., and Shih, C. W. 1995. "Computer simulation of impression creep using the hyperbolic sine stress law." *Materials Science and Engineering A*, 201(1-2), 50-57.
- Yang, F., Li, J. C., and Shih, C. W. 1995. "Computer simulation of impression creep using the hyperbolic sine stress law." Volume 201, Issues 1-2, October 1995, Pages , 201(1-2), 50-57.

ESTIMATION OF THE DYNAMIC EFFECT IN THE LIFTING OPERATIONS OF A BOOM CRANE

Luigi Solazzi, Giovanni Incerti and Candida Petrogalli
Department of Mechanical and Industrial Engineering

University of Brescia

Via Branze 38, 25123 Brescia, Italy

Email: luigi.solazzi@unibs.it, giovanni.incerti@unibs.it, candida.petrogalli@unibs.it

KEYWORDS

Lifting devices, overloading condition, vibration, motion command.

ABSTRACT

This paper describes a model for the study of the dynamic behavior of lifting equipments. The model here proposed allows to evaluate the fluctuations of the load arising from the elasticity of the rope and from the type of the motion command imposed by the winch. A calculation software was developed in order to determine the actual acceleration of the lifted mass and the dynamic overload inside the rope during the lifting phase. In the final part of the article an application example is presented, with the aim of showing the correlation between the magnitude of the stress and the type of the employed motion command.

INTRODUCTION

As is well known, the application of a load to a structure generates vibratory effects on the structure itself. This aspect is of a general nature and obviously it does not consider the type of structure and the rules for the load application, aspect, the latter, that influences the magnitude of the dynamic instead.

In a lifting equipment (cranes, bridge cranes, mobile cranes, etc..) this phenomenon appears to be particularly important, since it generates actions whose wrong assessment could compromise the structure, in particular as regards the phenomena of maximum stress, buckling, fatigue and equilibrium of the structure itself. Large bibliography of this subject has been elaborated in recent years (Bao, Zhang and Zhu 2011), as the problem, albeit overlooked in past years, now plays an important role during the design stage of a lifting device. To confirm what we said just think that the real force acting on the structure as a result of lifting can be equal to $1.5 \div 2$ times the nominal force generated from the hoist.

With the need to reduce the weight and increase the performances (see for example the mobile cranes where the frame is also more lightweight), this load increment

cannot be absolutely neglected (Matteazzi and Solazzi 2000) (Matteazzi and Solazzi 2005).

It is important to observe that the value of the dynamic effect is about equal or even greater to the value of the safety factor (adopting a deterministic design methodology) and therefore its wrong estimation may lead to the design of a non-safe device. Another fundamental aspect resides in the fact that the dynamic effect increases not only the action, but also the maximum number of stress cycles to which the device is subjected; as is well known, such cycles, while being of limited amplitude, contribute significantly to the fatigue life of the machine.

The basic parameters of the stress cycle as the amplitude (related to the intensity of the dynamic effect), the numerosity (related to the damping of the structure) and the temporal sequence (related to the sequence operations of lifting and handling of cargo) are variables which cannot be ignored (Solazzi 2004). In the face of these problems, regulatory agencies have proposed proper coefficients that, multiplying the load to be moved, allow the designer to consider these phenomena in the calculation.

Starting from the studies reported in the literature and from the analysis of the Standards for the design of lifting equipments, we developed a mathematical model that allows to estimate the load fluctuations caused by the elasticity of the rope and by the elasticity of the structure, in order to evaluate or estimate the dynamic overload inside the rope during lifting.

A BRIEF OVERVIEW OF INTERNATIONAL STANDARDS ABOUT LIFTING DEVICES

From the regulatory point of view, there are several Standards for the design of a lifting device. All introduce a coefficient of dynamic overload that depends on the class of the structure. This class is defined in relation to the number of stress cycles to which the structure is subjected, to the spectrum of the load and to the lifting speed. Figure 1 shows an example of the dynamical effect, evaluated by experimental test on a working platform (Solazzi 2009; Solazzi and Scalmana 2012). In general is possible to observe that at the beginning or at the end of the movement, the structure is subjected to oscillations and thus to accelerations that increase the load acting on

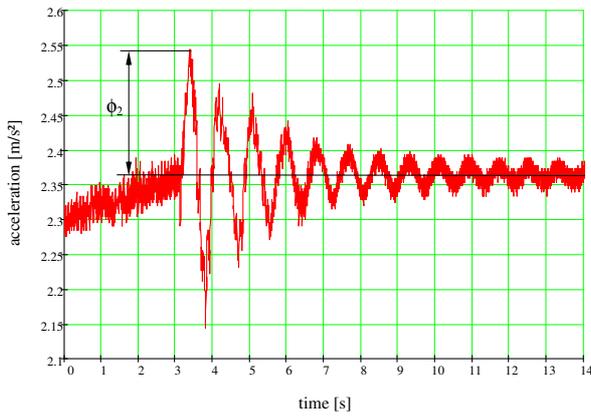


Figure 1: Dynamic overloads on a working platform in correspondence with a stoppage of the lifting motion.

the structure itself. Table 1 shows some values defined by these Standards, where β_2 is a parameter that depends on the hoisting class, V_0 is the velocity of lifting and ϕ_2 is the dynamical overloading.

Table 1: Dynamic effect as a function of the Standard.

UNI 9309 (1988) $\phi_2 = \phi_{2 \min} + \beta_2(V_0 - 0.2)$ $V_0 > 0.2 \text{ m/s}$		
Hoisting Class (β_2)	$\phi_{2 \min}$	$\phi_{2 \max}$
0.2	1.00	1.30
0.4	1.05	1.60
0.6	1.10	1.90
0.8	1.15	2.20
DIN 15017 (1975) $\phi_2 = \phi_{2 \min} + \beta_2 V_0$		
Hoisting Class (β_2)	$\phi_{2 \min}$	$\phi_{2 \max}$
0.132	1.00	1.30
0.264	1.20	1.60
0.396	1.30	1.90
0.528	1.40	2.20
UNI EN 13001 (2005) $\phi_2 = \phi_{2 \min} + \beta_2 V_0$		
Hoisting Class (β_2)	$\phi_{2 \min}$	
0.17	1.00	
0.34	1.05	
0.51	1.10	
0.68	1.15	

For the design and testing of lifting equipments the coefficients associated to the dynamic effect are in some cases higher than the normal safety factors defined by the Standards. Moreover these values of ϕ_2 do not depend neither on the stiffness of the structure nor on the type of motion law, aspects that can strongly affect the amount of overload.

MATHEMATICAL MODEL

In most cases of practical interest, elasto-dynamic models with one or two degrees of freedom (DOF) allow to simulate with good approximation the dynamic behavior of lifting devices. The use of a 1-DOF model can be justified in the case where the structure is very rigid. If the structure presents a stiffness comparable to the lifting system, as in the example presented in this work, a 2-DOF modelling is recommended (Matteazzi and Minoia 2007).

Figure 2 shows in schematic form a model with two degrees of freedom that can be used to analyze the dynamics of a lifting device.

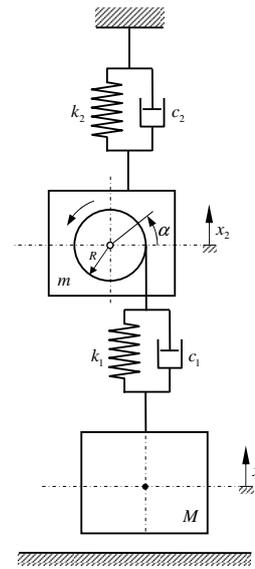


Figure 2: Two degrees of freedom model of a lifting device.

The vertical displacement of the load is represented by the coordinate x_1 , while the vertical displacement of the lifting structure is indicated by the coordinate x_2 . The symbols M and m respectively indicate the mass of the load and the mass of the lifting structure. The stiffness and the damping constant of the lifting system (ropes) are indicated by the symbols k_1 and c_1 ; in a similar way, the symbols k_2 and c_2 indicate the values of stiffness and damping of the structure. The rope is wound on a drum of radius R , driven by a geared motor unit; the model proposed here assumed to know the drive law of the drum i.e. the function $\alpha = \alpha(t)$. Figure 3 are represented the free body diagrams of the two masses, in order to highlight the forces that are generated during the lifting phase. In the picture the forces due to gravity were not represented because such forces are already balanced by the elastic reactions due to static deformations of the system. With the sign conventions shown Figure 3 we can write the following system of equations:

$$\begin{cases} M\ddot{x}_1 + c_1(\dot{x}_1 - \dot{x}_2) + k_1(x_1 - x_2) = R(k_1\alpha + c_1\dot{\alpha}) \\ M\ddot{x}_1 + m\ddot{x}_2 + c_2\dot{x}_2 + k_2x_2 = 0 \end{cases}$$

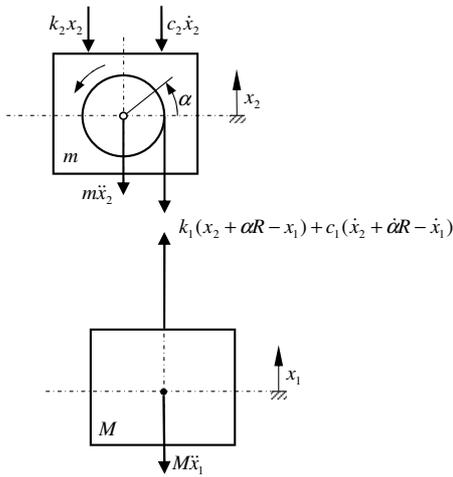


Figure 3: Free-body diagrams.

The above equations have been implemented in a software (Mathcad) and were solved numerically by going to determine variables x_1 and x_2 for example described in the following paragraph. As stated, the model here described requires to know the function $\alpha = \alpha(t)$ that defines the motion of the drum during the lifting phase. From a practical point of view, this motion law may be experimentally measured by mounting an angular displacement or a velocity transducer (encoder or tacho) on the axis of the drum.

Bearing in mind that, in most cases, the three-phase induction motor that drives the winch is controlled by a frequency converter, a device that allows you to adjust with a good approximation the acceleration and deceleration time intervals, it seems reasonable to adopt a law of motion with acceleration and deceleration that is easily definable in analytical form.

Therefore, for the purposes of the simulation of the dynamic behaviour of the system, we can assume valid the diagram represented in Figure 4, which can be immediately determined based on the knowledge of four parameters: the maximum speed $\dot{\alpha}_{max}$ and the three time intervals t_1 , t_2 and t_3 , which respectively represent the time duration of the phases of acceleration, constant speed and deceleration.

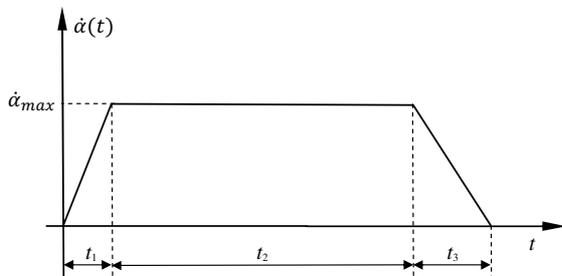


Figure 4: Angular velocity of the drum versus time.

APPLICATION OF THE MODEL: A BOOM CRANE

This section shows an example of an application of that previously reported. We have considered a boom crane with a load capacity of 50 tons and a reach of 80 m, designed and engineered according to the different dedicated Standards. It is the typical crane used in ports for loading and unloading of containers from ships. We made solid model and then the FEM model of the structure with the purpose, on one hand, to verify the crane and secondly to determine the displacement and its stiffness in different load configurations. Figures 5 and seq. report the design of the structure and its maximum displacement in the different load configurations typical for this kind of cranes. In particular, Figure 5 shows the displacements in static configuration (displacements induced only by own weight), Figure 6 shows the total and vertical displacements in the case where the load is positioned on the seaside on the ship and Figure 7 shows displacements in the case in which the load is positioned on the quay side.

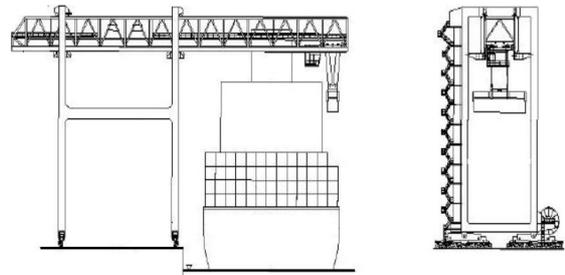


Figure 5: The retractable boom crane studied: boom length: 80 m, seaward side: 38 m, gantry width: 27 m, height (under hook): 45 m.

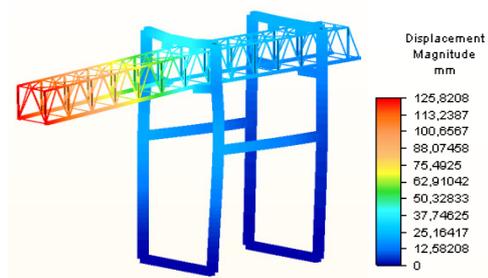


Figure 6: Structure displacement due to gravity effect.

The values of displacement and stiffness found are summarized in Table 2.

Table 2: Displacements and stiffness of the structure.

Load position	Displacement [mm]	Stiffness k_2 [N/mm]
Sea side	143.2	4190
Quay side	34.4	17450

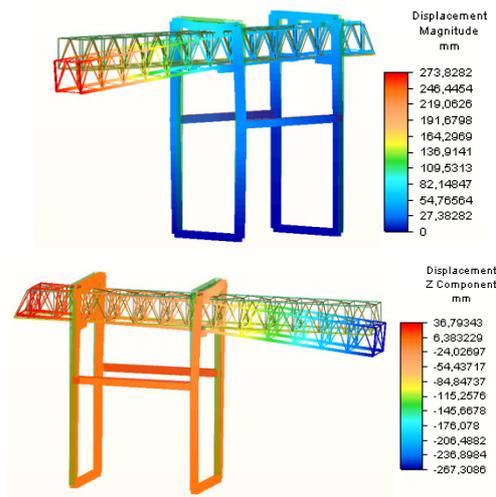


Figure 7: Structure displacement: load lifting from the seaward side: a) total displacement; b) vertical displacement.

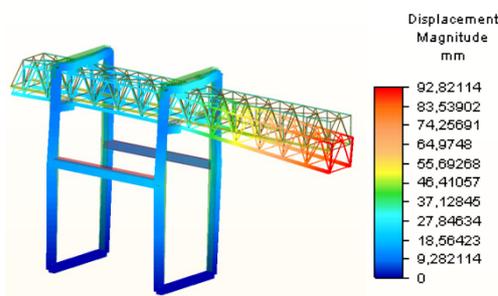


Figure 8: Structure displacement: load lifting from dock-side.

These values correspond also to a minimum and a maximum stiffness configuration for the structure.

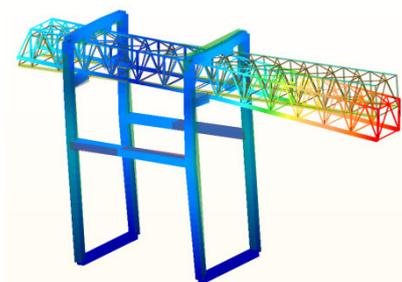


Figure 9: Trend of displacement of the crane relative to the first mode of vibration in vertical direction.

In parallel to the structural design a lifting system composed of a block and some pulleys has been employed.

In particular, we chose a system made by four pulleys (8 branches of rope 24 mm in diameter) and one made by five pulleys (10 branches rope 22 mm in diameter). On the basis of cables commercially available (in particular Warrington Seale type), we calculated both the stiffness

of each section that composes the lifting system and the package consisting of all the branches.

During development, we considered two different manufacturers of ropes, but we found very similar results in terms of stiffness. Because in practice, in the specific case of the container of 50 tons, the operation of lifting the load takes place at a variable height in relation to the position of the container that has to be moved on the ship, we performed a procedure in order to calculate the equivalent stiffness of the structure in relation both to the system lifting (8 or 10 branches) and to the height of the load (see Table 3).

Table 3: Displacements and stiffness of lifting system related to system configuration.

Ropes Nr.	Rope diam. [mm]	Lifting height [mm]	Displacement [mm]	Single rope Stiffness [N/mm]	Total Stiffness k_1 [N/mm]
8	24	65000	159.6	470	3760
8	24	45000	110.5	680	5430
8	24	10000	24.6	3050	24430
10	22	65000	152.0	400	3950
10	22	45000	105.2	570	5700
10	22	10000	23.4	2570	25660

To determine the mass of the lifting system involved in vibration, we performed a modal analysis considering the lifting of the load on both the sea and the port side and we calculated the natural frequencies of the system together with their mass participation. Figure 8 shows the displacement of the crane relative to the first mode of vibration. The first frequency found was about 1.5 Hz and. Being this value very low, we can already observe that the dynamic actions induced for example by an earthquake have a limited importance in structural design and verification of the crane (Solazzi 2011; Solazzi 2012). The value of the modal mass found for that configuration is equal to 64 tons, approximately 21% of the entire mass of the structure. By adding this value to that of the mass of the lifting device we determined the mass m necessary for the mathematical model. This mass resulted equal to 80 tons. On the base of the FEM results obtained, we performed a dynamic analysis using the model defined above. We chose a configuration such that the lifting speed was set on about 0.25 m/s and the time needed to get the speed regime condition was set on 4 sec.

The masses M and m , defined and calculated, were equal to $M = 50$ tons and $m = 80$ tons. The constants c_1 and c_2 have been fixed on 5×10^4 Ns/m.

Note that, in general, the damping values of c_1 and c_2 are strongly variable in relation to the type of structure (welded or bolted) and also to the type of rope (internal damping). In this work they were derived from the results of experimental tests by using the logarithmic decrement methodology (see Figure 1). The following graphs show the results in terms of speed and acceleration of the load and the structure (crane + lifting system).

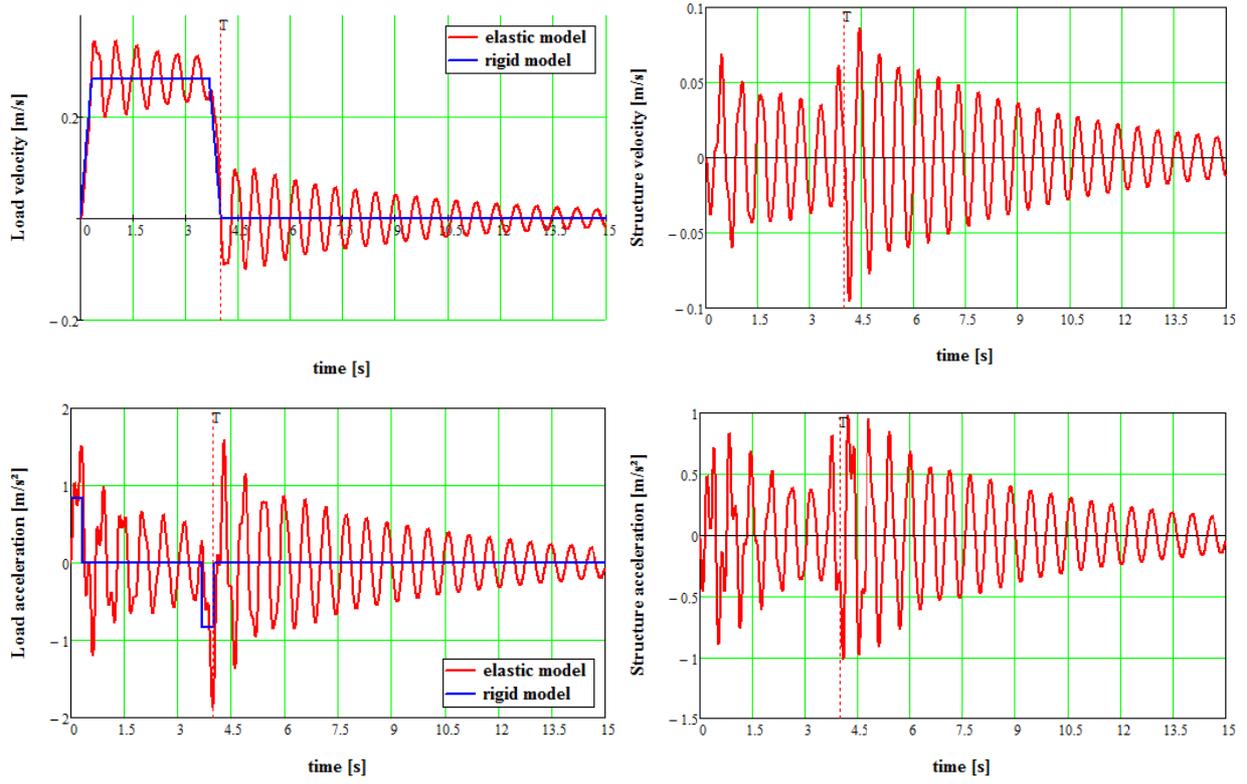


Figure 10: Load and structure velocity and acceleration in maximum stiffness condition.

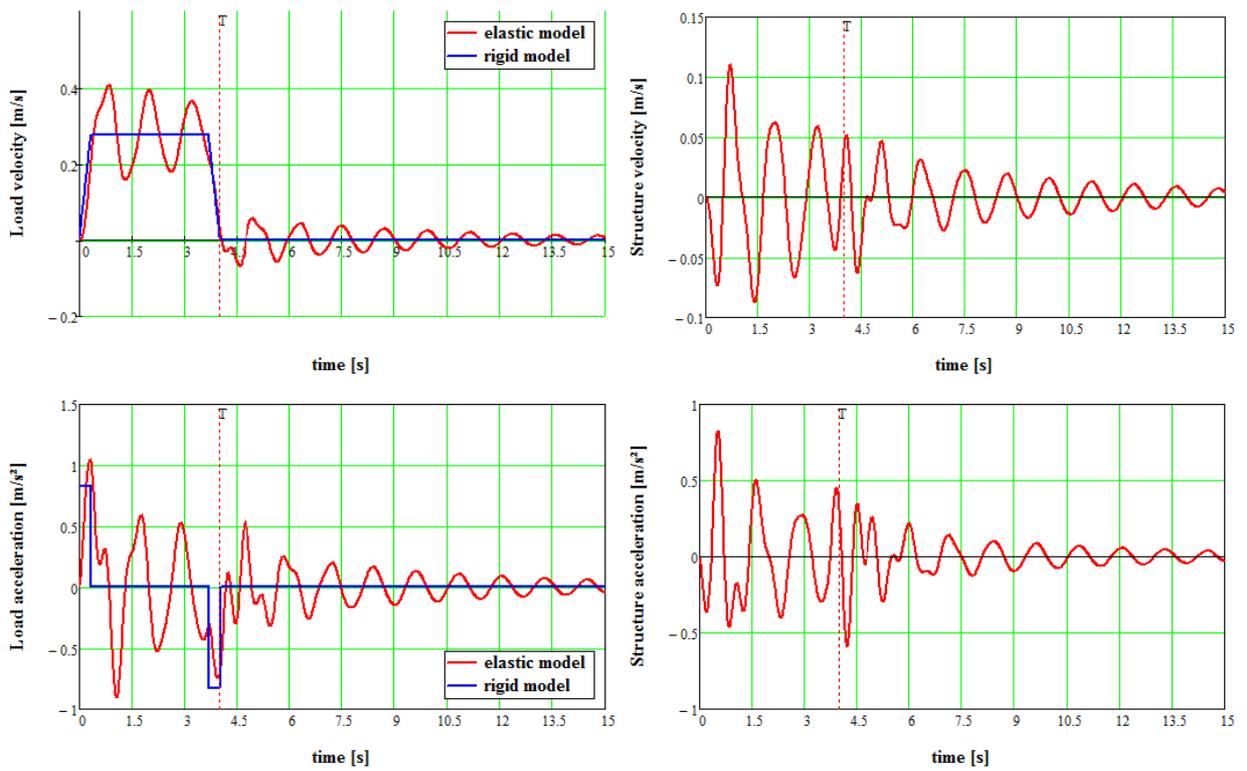


Figure 11: Load and structure velocity and acceleration in minimum stiffness condition.

Among the configurations of stiffness available we chose the two most significant and useful to make consideration about the dynamic behavior of the crane:

- the condition of maximum stiffness for the structure and for the lifting device ($k_1 = 25660 \text{ N/mm}$; $k_2 = 17450 \text{ N/mm}$, Figure 10);
- the condition of minimum stiffness for the structure and lifting device ($k_1 = 3760 \text{ N/mm}$; $k_2 = 4190 \text{ N/mm}$, Figure 11).

Comparing these graphs with experimental data it is possible not only to calibrate the model, but also to derive useful information for the crane design.

The natural frequencies of vibration of the structure are related to the stiffness and both the masses of the system with two degrees of freedom; this aspect also has important implications for the phenomenon of fatigue for the materials.

Since the purpose of this work is to evaluate the effect of overstressing induced by the operation of loading we also present in Figure 12) two plots that put in relation the acceleration of the load and the structure as a function of the ratio M/m and of the ratio k_1/k_2 . In particular for the Figure 12a the configuration is: $M = 50 \text{ tons}$, $m = \text{variable}$, $k_1 = 25660 \text{ N/mm}$, $k_2 = 17450 \text{ N/mm}$ and for the Figure 12b the configura-

tion is: $M = 50 \text{ tons}$, $m = 80 \text{ tons}$, $k_1 = 25660 \text{ N/mm}$, $k_2 = \text{variable}$.

As can be seen in more than one configuration the values found are higher than those considered in the Standards. An incorrect account of the same therefore could mean an higher risk for the structure both in terms of local or global buckling and in term of overturning.

CONCLUSIONS

An analysis of the dynamic effects in lifting devices has been presented in the paper. After an initial overview of the Standards in this field, we developed a lumped-parameter model with two degrees of freedom and we have applied it to the design of a boom crane, for quantifying the dynamic response. Through FEM simulations, we determined the stiffness of the structure in different geometric configurations. By the design of the lifting system (composed by ropes and drum) we evaluated the relation between stiffness and type of rope and number of branches. These parameters have been used and implemented in the calculation model and we determined the dynamic effect of the lifting of the load on the entire structure. We found that this effect, absolutely not negligible as regards the strength of the material and the crane stability, is also strictly dependent from the combination of structure and lifting system and from the mode of operation of the load. This procedure allows to estimate and compare in a very simple way all possible configurations of stiffness, mass, load velocity, etc. and to choose the best for a precise application. However, given the importance of this phenomenon, this work is the first phase of a larger project that will see the application of the model to other lifting equipment and its optimization through other rules of load movement.

REFERENCES

- Bao J., Zhang P., Zhu C. 2011. "Modeling and control of longitudinal vibration on flexible hoisting systems with time-varying length", *Procedia Engineering*, Vol. 15, 4521-4526.
- Matteazzi S., Minoia F. 2007. "Dynamical overloading of lifting appliances submitted to vertical movements: use of one degree of freedom oscillating systems, equivalent to two degree of freedom systems", *Int. J. Materials and Product Technology*, Vol. 30, 141-171.
- Matteazzi S., Solazzi L. 2000. "Determinazione dello stato di sforzo presente nelle ali di travi ad I di gru a ponte monorotaia", *Atti del XXIX Convegno Nazionale AIAS*, Lucca, 6-9 Settembre 2000.
- Matteazzi S., Solazzi L. 2005. "Influence of sequence of stress on fatigue damage", *Proceedings of the Ninth International Conference on Structural Safety and Reliability - ICOSSAR'05*, Rome, Italy, June 19-23, 2005.

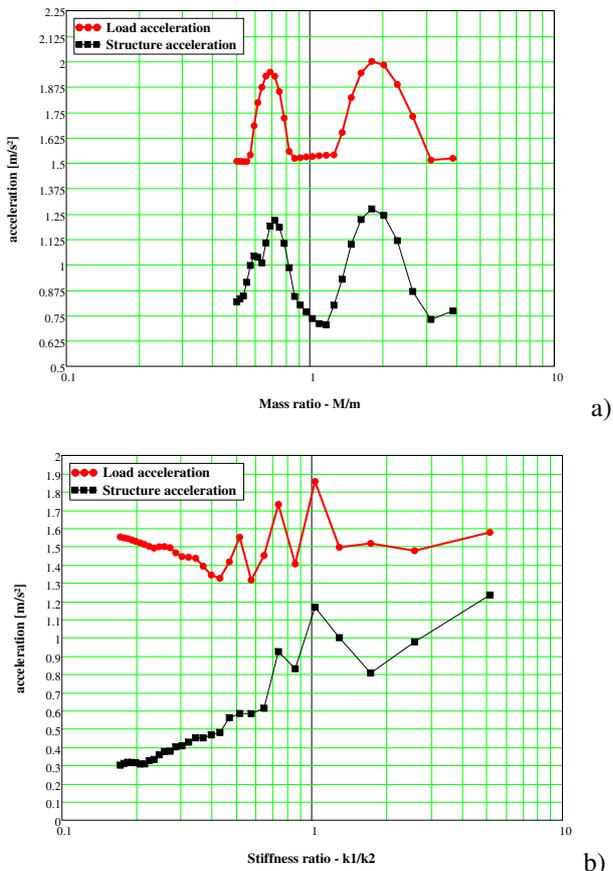


Figure 12: Load and structure acceleration vs M/m (a) and vs k_1/k_2 (b).

Solazzi L., 2004. “Sugli effetti dinamici indotti da differenti leggi di applicazione del carico negli apparecchi di sollevamento”, *Brescia Ricerche*, n°47 Giugno 2004 Pag. 18-22.

Solazzi L., 2009. “Design of an elevating work platform considering both the first and the second order effects and the inertial ones”, *2nd International Conference on Material and Component Performance under variable Amplitude Loading*, Darmstadt, Germany, March 23-26/2009.

Solazzi L. 2011. “Ship to shore crane subject to earthquake”, *11th International Conference on the Mechanical Behaviour of Materials*, June 5-9/2011 Milano, Italy.

Solazzi L., Scalmana, R. 2012. “New Design Concept for a Lifting Platform made of Composite Material”, *Applied Composite Materials*, An International Journal for Science and Application of Composite Materials, ISSN 0929-189X Appl Compos Mater DOI 10.1007/s10443-012-9287-2, Published online: 14 September 2012.

Solazzi L., 2012. “Dynamical behaviour of cranes subjected to different types of earthquake”, *XX International conference on Material Handling, Constructions and Logistics*, Belgrade, Serbia, 3-5 October 2012.

AUTHOR BIOGRAPHIES

LUIGI SOLAZZI was born in Montichiari, Italy and went to the University of Brescia, where he studied mechanical engineering and obtained his degree in 1994. He is obtained PhD in the 2003. He works in the Department of Mechanical and Industrial Engineering of Brescia University and works for Machine Design Group. His e-mail address is: luigi.solazzi@unibs.it.

GIOVANNI INCERTI graduated in Mechanical Engineering in 1990 at the University of Brescia, Italy. At the same university he received the Ph.D in Applied Mechanics in 1995. Actually he is associate professor at the University of Brescia and his teaching activity is related to the courses of Mechanical Vibrations, Applied Mechanics and Vibration Control. He is author of papers dealing with the dynamic analysis of cam systems, the mathematical modeling of devices for industrial automation, the mechanism design by optimization techniques and the study of robots and servomechanisms. His e-mail address is: giovanni.incerti@unibs.it.

CANDIDA PETROGALLI was born in Leno, Italy and went to the University of Brescia, where she studied mechanical engineering and obtained her degree in 2006. She is a PhD Student in Technologies and Energetic Systems for Mechanical Industry at the Department of Mechanical and Industrial Engineering of Brescia University and works for Machine Design Group. Her e-mail address is: candida.petrogalli@unibs.it.

Simulation of Intelligent Systems

TOWARDS EVOLUTIONARY DEEP NEURAL NETWORKS

Tomás H. Maul
Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih,
Malaysia
E-mail: Tomas.Maul@nottingham.edu.my

Chong Siang Yew
Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih,
Malaysia
E-mail: siang-yew.chong@nottingham.edu.my

Andrzej Bargiela
Computer Science
The University of Nottingham
Jubilee Campus, Wollaton Road, Nottingham
UK
E-mail: Andrzej.Bargiela@nottingham.ac.uk

Abdullahi S. Adamu
Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih,
Malaysia
E-mail: khyx1asa@nottingham.edu.my

KEYWORDS

Evolutionary artificial neural networks, neuroevolution, deep neural networks, hybrid neural networks.

ABSTRACT

This paper is concerned with the problem of optimizing deep neural networks with diverse transfer functions using evolutionary methods. Standard evolutionary (SEDeeNN) and cooperative coevolutionary methods (CoDeeNN) were applied to three different architectures characterized by different constraints on neural diversity. It was found that (1) SEDeeNN (but not CoDeeNN) changes parameters uniformly across all layers, (2) both evolutionary approaches can exhibit good convergence and generalization properties, and (3) increased neural diversity improves both convergence and generalization. In addition to clarifying the feasibility of evolutionary deep neural networks, we suggest a guiding principle for synergizing evolutionary and error gradient based approaches through layer-change analysis.

INTRODUCTION

This paper attempts to bring together two active but mostly independent research areas within artificial neural networks (ANN), i.e., deep neural networks (DNN) (Hinton et al. 2006) and evolutionary neural networks (ENN) (Yao 1999) (also referred to as neuroevolution). To the best of the authors' knowledge, the literature does not yet show explicit signs of systematic work on evolutionary deep neural networks (EDeeNN). We attempt to take a few early steps in this direction by exploring different evolutionary approaches and different architectural constraints, with the long-term goal of finding efficient and accurate EDeeNNs. We do this in the context of hybrid neural networks (Gutiérrez et al. 2009; Gutiérrez et al. 2011), particularly neural diversity machines (NDM) (Maul 2013). We also investigate how different layers change throughout the optimization process, in order to

understand possible underlying causes behind different convergence speeds and generalization capabilities.

The following section gives an overview of the related work, which includes evolutionary and deep neural networks. Subsequently the four main hypotheses underlying the work are summarized, followed by a methodology section which covers neural network architectures, evolutionary algorithms and the experimental setup adopted. After this, the main results are summarized, followed by a discussion and several conclusions.

RELATED WORK

Briefly put, ENN are neural networks whose weights and possibly architectures are tuned primarily via global stochastic optimization algorithms (e.g. genetic algorithms). Early classic references include (Belew et al. 1991) and (Whitley et al. 1990). Recent examples include (Gutiérrez et al. 2011) and (Gauci and Stanley 2010). For useful reviews refer to (Yao 1999) and (Floreano et al. 2008). Although difficulties in scaling ENNs to large models and/or data-sets have prevented this approach from becoming mainstream, ENNs offer several advantages such as the ability to optimize novel types of ANN for which learning algorithms cannot easily be derived and the potential to fulfill the promise of natural computation by effectively incorporating further biological elements into ANN (e.g. evolving developmental and learning rules in modular neural networks).

Deep neural networks (DNNs), which essentially consist of ANN with “many” layers (e.g. four or more connection layers) are not a new concept. They have been studied for many years in models such as the Neocognitron (Fukushima 1980) and Convolutional Neural Networks (LeCun and Bengio 1995). However, it was only after 2006 with the publication of (Hinton et al. 2006), and the development of methods that allowed for the efficient and consistent training of more flexible DNNs, that they finally became a popular topic.

Subsequently, many papers have been written and several major machine learning or pattern recognition competitions have been won, based on DNN breakthroughs. For a recent example, refer to the winner of the “Segmentation of neuronal structures in EM stacks challenge - ISBI 2012”, which used a DNN with convolutional and max-pooling layers (Ciresan et al. 2012). For a useful review of deep neural networks refer to (Bengio 2009). One of the main reasons why DNNs are so attractive consists of the fact that they consistently form good representations of the underlying causes of the data (automated feature construction) that lead to good generalization properties.

As mentioned, the literature does not yet seem to show any explicit signs of the combination of these two areas. Implicitly EDeeNNs have been studied in NEAT (Stanley and Miikkulainen 2002) where networks can in principle evolve to significant depths, and Neural Diversity Machines (NDM) (Maul 2013), where recurrence can be seen as depth in time. However, explicit studies as reported in this paper still seem to be lacking. More specifically, we address the question of the relative merits of two different evolutionary paradigms: (1) standard evolutionary Deep Neural Networks (SEDeeNN) and (2) cooperative coevolutionary Deep Neural Networks (CoDeeNN), and three different types of architectural constraints: (1) any type of transfer function (TF) for each neuron, (2) any type of TF per layer and (3) any type of TF for the whole network, where a TF consists of a particular combination of a weight function (WF), also commonly referred to as activation function, and a node function (NF), also commonly referred to as output function.

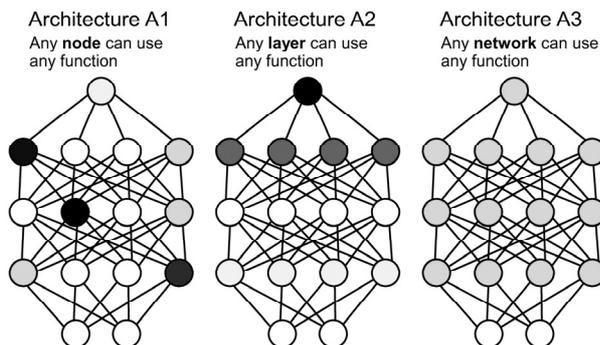


Figure 1: Three types of neural diversity architectures

HYPOTHESES

The experimental study was guided by four simple hypotheses: (1) in SEDeeNN the amount of change is proportional to layer depth (i.e. deeper (closer to the output) layers change more) whereas in CoDeeNN the amount of change is more balanced across layers, (2) convergence is faster for CoDeeNN compared to SEDeeNN, (3) CoDeeNNs generalize better (have lower test errors on average) than SEDeeNNs and (4) unconstrained architectures converge faster but do not

necessarily lead to the best generalization properties. Interestingly our results contradicted most of these hypotheses and revealed further insights and questions, such as the fact that the evolutionary and traditional error gradient approaches appear to be somewhat complementary in terms of how layers change and the question of how best to capitalize on this complementarity.

METHODS

Neural Architectures

The general neural architecture adopted in this paper is based on the NDMs reported in (Maul 2013), which are hybrid ANNs with constraints on the minimal amount of TF diversity allowed. The main difference here is that the connectivity has been restricted to being feedforward, with unrestricted depth or number of layers. This special case of NDMs can be referred to as feedforward NDMs. The main experiment in this paper used networks with four connection layers (i.e. three layers of hidden units), which is minimally but sufficiently deep, where each layer of hidden-units consisted of four nodes.

Several sub-architectures were allowed where each one was characterized by a different degree of freedom with regards to how TF diversity was used. Refer to Figure 1 for a diagrammatic representation of these three architectures. In architecture A1 any TF can be selected for any node, whereas in A2 any TF can be selected for any layer, and in A3 any TF can be selected for the whole network. Thus, only the first two conditions that define an NDM (Maul 2013) are applicable to these restricted (feedforward) NDMs: (1) at least 3 WFs and 3 NFs must be available for nodes and (2) nodes can exhibit any combination of the specified weight and node functions (corresponding to a minimum of 9 TFs).

Table 1 summarizes the weight and node functions used, where a_j refers to the WF output of node j , x_i refers the activation value of inputting node i , w_{ji} refers to the weight of the connection from node i to node j , c refers to some constant and o_j refers to the NF output of node j

Evolutionary Approaches

Two evolutionary approaches were tested, i.e.: (1) standard evolution and (2) cooperative coevolution. In the first case (SEDeeNN), solution vectors were treated as a whole (with no separate treatment for individual layers) and four different global heuristics (variation operators) were employed, i.e.: mutation, cross-over, differential evolution and probabilistic mingling (our implementation of rank-based uniform crossover (Semenkin and Semenkina 2012) and (Ackley 1987)). All heuristics except for the last one are commonly found in the literature. Probabilistic mingling is a form of cross-over where a pair of solutions (e.g. solutions s_1

and s2) is scanned parameter by parameter, in order to create a third solution (e.g. solution s3). If s1 is fitter than s2 then the probability of selecting a parameter from s1 (for inclusion in s3) is 0.75 as opposed to 0.25 for s2, and conversely if s2 is fitter. Population dynamics was based on initialization, expansion of the initial set using global heuristics, trimming by an elitist and diversity-preserving selection method and possible padding with random solutions (refer to Algorithms 1 and 2 for high-level overviews of the different optimization approaches).

Table 1: NDM weight and node functions

Type	Functions	Equations
WF	Inner product	$a_j = b_j + \sum_{i=1}^n w_j x_i$
WF	Euclidean distance	$a_j = \sqrt{\sum_{i=1}^n (w_j - x_i)^2}$
WF	Product	$a_j = \prod_{i=1}^n c w_j x_i$
WF	Higher-order product	$a_j = \prod_{i=1}^n x_i^{w_j} $
WF	Standard deviation	$a_j = \left(\sum_{i=1}^n w_j \right) \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$
WF	Minimum	$a_j = \min(w_j, x_1 \cdots w_j, x_n)$
WF	Maximum	$a_j = \max(w_j, x_1 \cdots w_j, x_n)$
NF	Identity	$o_j = a_j$
NF	Sigmoid	$o_j = \frac{1}{1 + e^{-a_j}}$
NF	Gauss. w/ thresh.	$o_j = e^{-\frac{(a_j)^2}{c}}$, if $o_j = d$ then $o_j = 1$

For the basic CoDeeNN, we reused as many aspects of SEDeeNN as possible (e.g. global heuristics and selection method) to facilitate comparisons. The main difference pertained to the creation of multiple sub-populations, one for each layer. The main challenge here consisted of the evaluation of each sub-solution in each sub-population. Note that each sub-solution represents a single layer and therefore cannot be evaluated independently. Evaluation was done in a similar way as with the Enforced SubPopulations method (Gomez and Miikkulainen 2003), i.e.: a sufficient number of whole solution vectors was created from random combinations of sub-solutions (one for each layer sub-population), which could then be evaluated in the normal way, and whereby the cost of a sub-solution could be subsequently computed by averaging the cost of all the networks it participated in.

Experimental Setup

The paper’s main experiment was designed to verify the four hypotheses outlined in the Hypotheses section. This experiment involved running 30 tests for every possible combination of architecture (i.e. A1, A2 and A3) and evolutionary approach (i.e. SEDeeNN and CoDeeNN), and storing information on training and test errors across 100 generations. The experiment was conducted using five different data-sets, i.e.: XOR, bar-circle, u-shape, Iris and Accute Inflammations (hereafter abbreviated as Inflammation). The bar-circle and u-shape datasets are

two simple and synthetic data-sets (depicted in Figure 2), whereas the last two datasets were obtained from the UCI Machine Learning Repository (Blake and Merz 1998). Apart from training and test errors, two other performance metrics were computed, i.e.: relative generalization capacity (RGC) and average layer-change. The former metric was computed for each combination of data-set, evolutionary approach and architecture, by dividing average training errors by average test errors for each generation, and then averaging all of these ratios. The latter was determined by computing the average parametric absolute change for each layer of the best solution (across 30 tests), generation by generation, and then averaging these changes (across 100 generations). For more information relating to algorithms and parameter settings please contact the corresponding author.

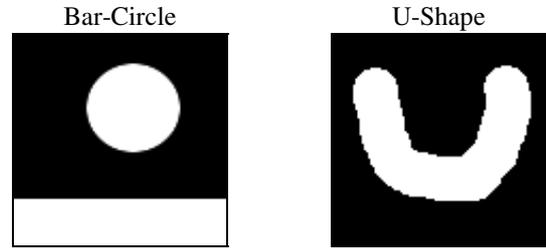


Figure 2. Example synthetic data-sets

RESULTS

Table 2 gives an overall summary of the results. The first column depicts for each data-set, a list of approaches ordered from best to worst in terms of average training error. This column is indicative of the convergence properties of different approaches. Several simple observations are possible here: (1) results for experiments involving architectures A1 and/or A2 consistently lead to faster convergence (seven cases for A1 and three cases for A2), (2) both SEDeeNN and CoDeeNN led to fastest convergences (in bold), (3) CA3 was consistently the worst (slowest) approach, (4) CoDeeNN was used in the best approaches for the more complex data-sets (i.e. Iris and Inflammation) whereas SEDeeNN was used in the best approaches for the simpler data-sets (i.e. XOR, bar-circle and u-shape), (5) the A2 architecture led to the fastest convergence for two data-sets (i.e. u-shape and Inflammation). Refer to the left-hand side of Figure 3 for the averaged training curves (over 30 tests) for the bar-circle, u-shape, Iris and Inflammation data-sets. The x-axes depict generations, whereas the y-axes depict classification errors. These results demonstrate the usefulness of neural diversity in evolutionary deep neural networks, both in terms of convergence and generalization. Refer to Figure 4 for training and test error curves averaged across data-sets and architectures.

It is important to note that although CoDeeNN converged the fastest for two cases (i.e. Iris and Inflammation), this convergence was measured in terms

of training error relative to generation and not actual computational time. In actuality, because of the difficulty and computational complexity of evaluating layers in a cooperative coevolutionary context, CoDeeNN is significantly slower than SEDeeNN.

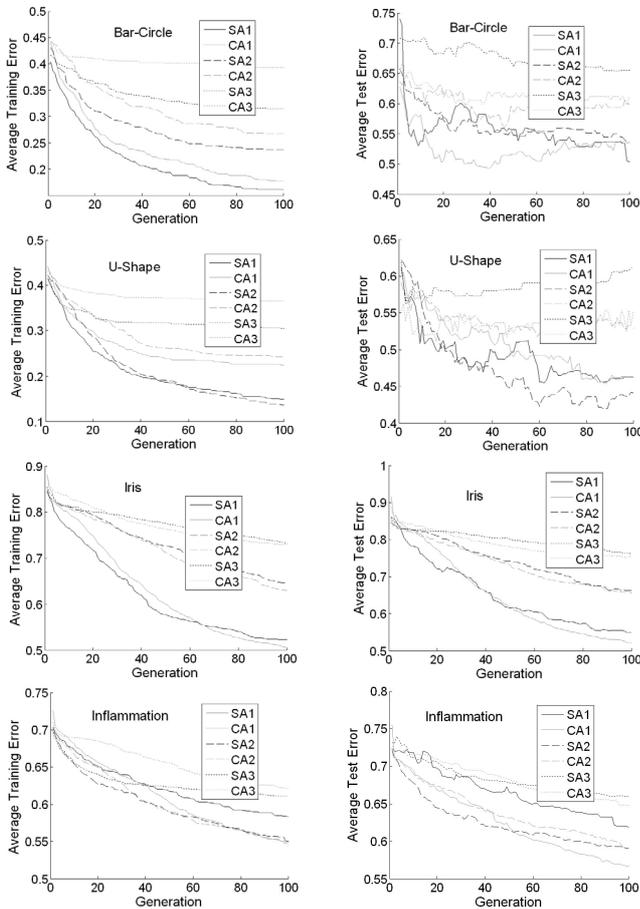


Figure 3. Training errors (left) and test errors (right)

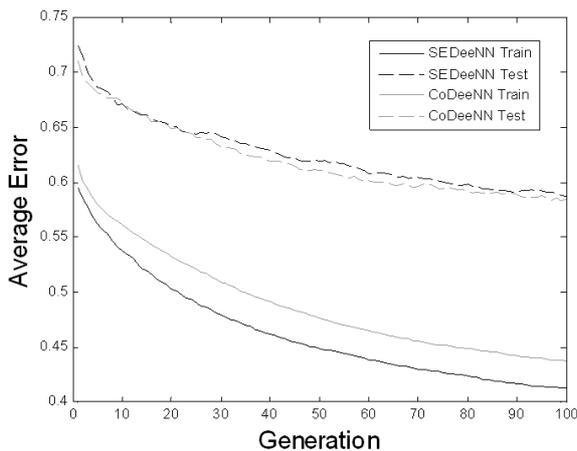


Figure 4: Average errors across bar-circle, u-shape, Iris and Inflammation

The second column in Table 2 is similar to the first column except that the ordering of approaches is done according to the average test error at generation 100. This column is indicative of the generalization

properties of different approaches. Because of the small size and nature of the XOR data-set we do not use it here to draw conclusions regarding generalization. As before, several simple observations can be made: (1) the generalization ranking of approaches (second column) is similar (although not equivalent) to the convergence ranking of approaches (first column), (2) the top-two generalizers always involve architectures A1 and/or A2 (six cases for A1 and two cases for A2), (3) the best generalizers may involve either SEDeeNN or CoDeeNN, (4) the worst generalizers always involve architecture A3 (as mentioned we exclude XOR), (5) the two worst generalizers were always CA3 and SA3 and the latter was consistently the worst one. The right-hand side of Figure 3 depicts average test error curves for the bar-circle, u-shape, Iris and Inflammation data-sets.

Algorithm 1: High-level overview of SEDeeNN

```

1: function SEDeeNN(data, p)
2:   noStop = true;
3:   trimSols = initializeSolutions(data, p);
4:   While noStop
5:     coSols = doCrossOver(trimSols, p);
6:     pmSols = doProbabilisticMingling(trimSols, p);
7:     mSols = doMutation(trimSols, p);
8:     deSols = doDifferentialEvolution(trimSols, p);
9:     sols = [trimSols; coSols; pmSols; mSols; deSols];
10:    sols = evaluateSolutions(sols, data, p);
11:    sols = sortSolutions(sols);
12:    trimSols = trim(sols, p);
13:    noStop = check stopping conditions;
14:  End

```

Algorithm 2: High-level overview of CoDeeNN

```

1: function CoDeeNN(data, p)
2:   noStop = true;
3:   pops = initializePopulations(p);
4:   While noStop
5:     pops = evaluateSolutionsCoCo(pops, data, p);
6:     For i = 1 to number of populations
7:       aPop = pops{i};
8:       coSols = doCrossOver(aPop, p);
9:       pmSols = doProbabilisticMingling(aPop, p);
10:      mSols = doMutation(aPop, p);
11:      deSols = doDifferentialEvolution(aPop, p);
12:      pops{i} = [coSols; pmSols; mSols; deSols];
13:    End
14:    noStop = check stopping conditions;
15:  End
16:  pops = evaluateSolutionsCoCo(pops, data, p);

```

The third column in Table 2 depicts those approaches which obtained the largest relative generalization capacity. The fact that CA3 obtained the best RGC score in three cases is most probably more a reflection of the fact that it was the poorest converger. However, what is more noteworthy is that for the Inflammation data-set, CA1 obtained the highest RGC score when it was also the best generalizer (second column). In the very least this provides more evidence that increased neural diversity (i.e. A1) does not necessarily impact generalization capacity in a negative way, either in an

absolute (i.e. average final test error) or relative (i.e. RGC) sense.

Columns four and five address the question of whether different layers change on average to different extents for different approaches. The most obvious observation here is that for CoDeeNN there is always differential layer-change, regardless of the architecture and data-set used, whereas for SEDeeNN, in an overwhelming majority of cases, layer-change is uniform (the two possible exceptions are “bar-circle SA3” and “Iris SA2”). Refer to Figure 5 for specific examples of the distinction between SEDeeNN and CoDeeNN in terms of layer-changes, using different data-sets. The y-axes depict the average amount of parametric change per layer, normalized by the number of parameters used by each layer. These examples show how SEDeeNN tends to be associated with uniform layer-changes, whereas CoDeeNN, where layers evolve semi-autonomously, is associated with different degrees of parametric change per layer. Legend: conL1 ... conL2 = connection layer 1 ... connection layer 4.

Considering the Inflammation case (Figure 5, bottom row) across architectures A1, A2 and A3, notice how the middle layers always exhibit maximal layer-change, whereas the last layer always exhibits minimal change, and the first layer varies. Furthermore notice how architecture A1 involves the least overall change. This is in contrast to layer-change dynamics in error gradient based approaches, where it is generally hard for error-based information to penetrate earlier layers, and therefore suggests a useful guiding principle for the integration of evolutionary and gradient based approaches.

Several sanity checks were also conducted using SEDeeNN with an A1 architecture adopting seven connection layers (node/layer architecture: [2 8 4 4 4 4 1]), on several simple data-sets including XOR and other 2D data-sets similar to those in Figure 2, whereby it was demonstrated that deeper networks could indeed converge to zero training error in less than 100 generations.

DISCUSSION

Hypotheses

The paper’s first hypothesis was completely disproven by the experimental results, which demonstrated that layer-change tends to be more consistent across layers for SEDeeNN, in contrast to CoDeeNN where the middle layers tend to change the most and the last layer tends to change the least (see Figure 5). Looking back at Table 2, one can see that 15 out of 15 CoDeeNN cases involve differential layer-change, whereas only 2 out of 15 SEDeeNN cases involve differential layer change. The table legend consists of: Train (100) = list of approaches, ordered from best to worst, in terms of

average training error at generation 100; Test (100) = list of approaches, ordered from best to worst, in terms of average test error at generation 100; SA1 ... CA3 = SEDeeNN Architecture 1 ... CoDeeNN Architecture 3; Max RGC = approach with largest relative generalization capacity; LC = layer-changes; s = no significant difference between layers in terms of layer-changes; D = significant difference between layers in terms of layer-changes.

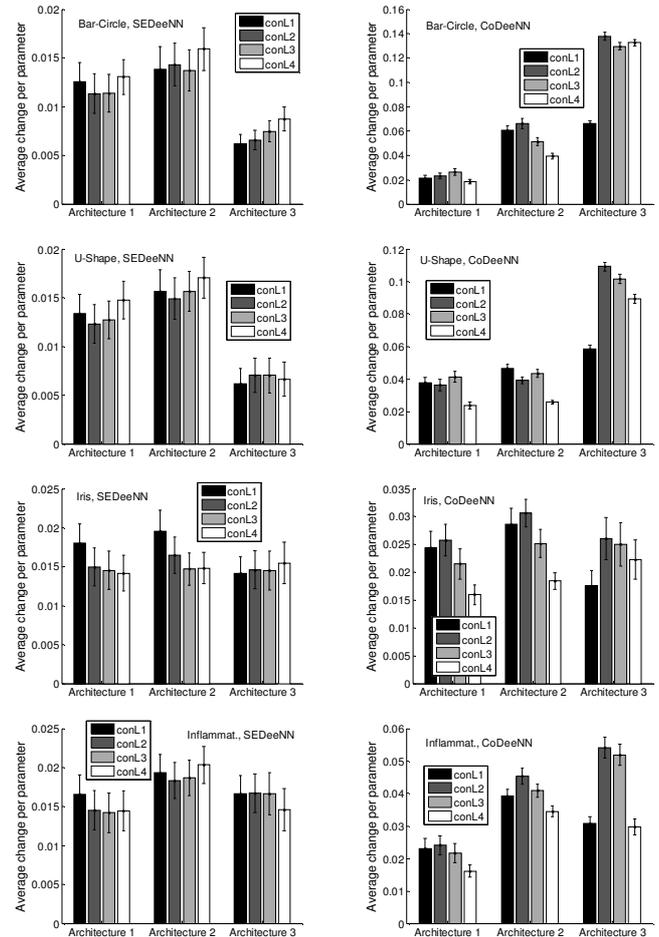


Figure 5: Layer changes for SEDeeNN and CoDeeNN

The second hypothesis was also disproven. Overall SEDeeNN seems to have better convergence properties (e.g. XOR, bar-circle and u-shape) although there are cases where CoDeeNN seems to be better (e.g. Iris and Inflammation). Going back to Table 2, it can be seen that 6 out of the 10 top-two (winner and runner-up) convergers for all data-sets use SEDeeNN whereas 4 out of 10 use CoDeeNN. Looking at training error curves averaged across datasets and architectures (Figure 4) the overall convergence advantage of SEDeeNN is clear. Note that this conclusion, so far, is only valid for a basic implementation of cooperative coevolution of layers in a deep neural network. It is possible that more advanced implementations may change this conclusion.

The third hypothesis was also disproven. Although there is definite evidence for the generalization capability of CoDeeNN (e.g. Inflammation and Iris), SEDeeNN is by no means inferior (e.g. bar-circle and u-shape). In fact, closer inspection of Table 2 reveals that 5 of the 8 top-two generalizers across all data-sets (excluding XOR) use SEDeeNN whereas only 3 out of 8 use CoDeeNN. Notice also how SA2 has the highest RGC score for Iris and how the SEDeeNN and CoDeeNN average test error curves in Figure 4 are virtually indistinguishable.

Table 2: Summary of results

Data	Train (100)	Test (100)	Max RGC	SEDeeNN LC			CoDeeNN LC		
				A1	A2	A3	A1	A2	A3
XOR	SA1, SA2, SA3, CA1, CA2, CA3	SA2, SA3, SA1, CA3, CA2, CA1	CA3	S	S	S	D	D	D
Bar-circle	SA1, CA1, SA2, CA2, SA3, CA3	SA1, CA1, SA2, CA2, CA3, SA3	CA3	S	S	D	D	D	D
U-shape	SA2, SA1, CA1, CA2, SA3, CA3	SA2, SA1, CA1, CA2, CA3, SA3	CA3	S	S	S	D	D	D
Iris	CA1, SA1, CA2, SA2, CA3, SA3	CA1, SA1, CA2, SA2, CA3, SA3	SA2	S	D	S	D	D	D
Inflammat.	CA2, CA1, SA2, SA1, SA3, CA3	CA1, SA2, CA2, SA1, CA3, SA3	CA1	S	S	S	D	D	D

The first part of the fourth and final hypothesis was confirmed. In general the more unconstrained the architecture the faster the convergence. A1 is the most unconstrained architecture and tends to be the most highly ranked one (average convergence ranking; not shown). Conversely A3 is the most constrained architecture and tends to be the lowest ranked one. Note that the u-shape and Inflammation data-sets provide interesting exceptions in that A2 converges faster than A1. Note how in Table 2, the second half of the hypothesis was mostly disproven, since in spite of unconstrained diversity allowing greater convergence speed this did not consistently affect generalization in a negative way. Unconstrained architectures not only tended to provide better convergence but also better generalization properties.

Neural Diversity

The fact that these results mostly disprove the original hypotheses is good news for neural diversity. Neural diversity was repeatedly shown to improve convergence speed without simultaneously jeopardizing generalization capacity. Note that the more unconstrained an architecture is, the more neural diversity it exhibits. The results show that neural diversity can be added to deep neural networks, leading to improved convergence and generalization. Moreover, this addition doesn't even demand a more sophisticated optimization approach such as cooperative coevolution in order to deal with the typical issues posed by deep layers. On the contrary, a standard evolutionary approach characterized by basic global heuristics such as mutation, cross-over, differential evolution and probabilistic mingling, was shown to be sufficiently effective.

Synergizing Evolution and Error Gradients

In general, one of the major problems with deep neural networks pertains to vanishing error gradients, which makes it difficult for error derivatives to percolate from outer to inner layers. In EDeeNN this is indirectly reflected in how much different layers change from generation to generation (this is the ENN version of the problem). Before running the experiments it was believed that a standard evolutionary approach would lead to more extensive changes at outer layers, with limited changes at inner layers, mirroring the error gradient case. Fortunately the results contradicted this expectation, demonstrating that SEDeeNN involves uniform layer changes whereas CoDeeNN involves relatively more changes in middle and inner layers. This suggests that evolutionary and error gradient based approaches can be complementary, and that the potential benefit of this synergy is likely to be more fully exploited if we are guided by a deeper understanding of layer-change within each approach and in the context of their integration.

Conclusions

This paper has shown that neural diversity tends to improve the convergence and generalization properties of small deep neural networks. Moreover, standard evolutionary algorithms appear to be at least as good as cooperative coevolution in optimizing deep neural networks with diverse transfer functions. Also, standard evolutionary methods were shown to be capable of changing parameters consistently across all layers. We believe that the significance of this work lies in the demonstration that neural diversity, standard evolutionary methods and the analysis of layer-changes are all fruitful priorities for future work in the area of evolutionary deep neural networks.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for helpful comments and the Faculty of Science, University of Nottingham Malaysia Campus, for its support.

REFERENCES

- Hinton, G.E., Osindero, S. and Teh, Y.W. 2006. "A fast learning algorithm for deep belief nets," *Neural computation*, 18(7), pp. 1527-1554.
- Yao, X. 1999. "Evolving artificial neural networks," *Proceedings of the IEEE*, 87(9), 1423-1447.
- Gutiérrez, P.A., Hervás, C., Carbonero, M. and Fernández, J.C. 2009. "Combined projection and kernel basis functions for classification in evolutionary neural networks," *Neurocomputing*, 72(13), 2731-2742.
- Gutiérrez, P.A. and Hervás-Martínez, C. 2011. "Hybrid artificial neural networks: models, algorithms and data". In *Advances in Computational Intelligence*, Springer Berlin Heidelberg, 177-184.
- Maul, T. 2013. "Early experiments with neural diversity machines," *Neurocomputing*, 113, 36-48.

- Belew, R.K., McInerney, J. and Schraudolph, N.N. 1991. "Evolving networks: Using the genetic algorithm with connectionist learning." In Proc. Second Conference on Artificial Life, 511-547.
- Whitley, D., Starkweather, T. and Bogart, C. 1990. "Genetic algorithms and neural networks: Optimizing connections and connectivity," *Parallel computing*, 14(3), 347-361.
- Gutiérrez, P.A., Hervás-Martínez, C. and Martínez-Estudillo, F.J. 2011. "Logistic regression by means of evolutionary radial basis function neural networks," *IEEE Transactions on Neural Networks*, 22(2), 246-263.
- Gauci, J. and Stanley, K.O. 2010. "Autonomous evolution of topographic regularities in artificial neural networks," *Neural computation*, 22(7), 1860-1898.
- Floreano, D., Dürr, P. and Mattiussi, C. 2008. "Neuroevolution: from architectures to learning," *Evolutionary Intelligence*, 1(1), 47-62.
- Fukushima, K. 1980. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, 36(4), 93-202.
- LeCun, Y. and Bengio, Y. 1995. "Convolutional networks for images, speech, and time series," In *The handbook of brain theory and neural networks*, Cambridge, MA: MIT Press.
- Ciresan, D., Giusti, A. and Schmidhuber, J. 2012. "Deep neural networks segment neuronal membranes in electron microscopy images," In *Advances in Neural Information Processing Systems 25*, 2852-2860.
- Bengio, Y. 2009. "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, 2(1), 1-127.
- Stanley, K.O. and Miikkulainen, R. 2002. "Evolving neural networks through augmenting topologies," *Evolutionary computation*, 10(2), 99-127.
- Gomez, F.J. and Miikkulainen, R. 2003. "Active guidance for a finless rocket using neuroevolution," In *Genetic and Evolutionary Computation—GECCO*, 2084-2095.
- Blake, C. and Merz, C.J. 1998. "UCI Repository of machine learning databases."
- Semenkin, E., and Semenkina, M. 2012. "Self-configuring genetic programming algorithm with modified uniform crossover". In *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-6.
- Ackley, D. 1987. "A Connectionist Machine for Genetic Hillclimbing", vol. 28 of *The Kluwer International Series in Engineering and Computer Science*, Kluwer Academic, Boston, Mass, USA, 1987.

AUTHOR BIOGRAPHIES

TOMAS H. MAUL was born in Madeira, Portugal and did a BSc. in Biological Psychology at the University of St. Andrews, an MSc. in Computer Science at Imperial College and a PhD. in Computational Neuroscience at the University of Malaya. He worked for two years at MIMOS Bhd. as a Senior Researcher in the fields of Pattern Recognition and Computer Vision. He is currently an Associate Professor at the University of Nottingham Malaysia Campus, where he conducts research in the areas of Neural Computation, Optimization and Computer Vision. His e-mail address is Tomas.Maul@nottingham.edu.my and his Web-page is <http://kcztm.jupiter.nottingham.edu.my/>

ANDRZEJ BARGIELA is Professor of Computer Science at the University of Nottingham. Until recently he was Director of Computer Science at the University of Nottingham, Malaysia Campus. He is a member of the Automated Scheduling and Planning research group in the School of Computer Science at the University of Nottingham. Since 1978 he has pursued research focused on processing of uncertainty in the context of modelling and simulation of various physical and engineering systems.

CHONG SIANG YEW received the B.Eng. (Hons.) and M.Eng.Sc. degrees in electronics engineering from Multimedia University, Melaka, Malaysia, in 2002 and 2004, respectively, and the Ph.D. degree in computer science from the University of Birmingham, Edgbaston, Birmingham, U.K., in 2007. He is currently an Honorary Research Fellow with the School of Computer Science, University of Birmingham. He was a Research Associate with the Centre of Excellence for Research in Computational Intelligence and Applications, School of Computer Science, University of Birmingham, in 2007. He joined the School of Computer Science, University of Nottingham, Semenyih, Malaysia, in 2008. He is currently a member of the Automated Scheduling, Optimization and Planning Research Group, School of Computer Science, University of Nottingham, U.K.

ABDULLAHI SHUAIBU ADAMU received the B.S.(Hons) degree in Computer Science from the University of Nottingham Malaysia Campus, Malaysia, in 2011 and is currently a PhD candidate at the same University within the School of Computer Science. Abdullahi Adamu is conducting research in the area of hybrid artificial neural networks, with special focus on the issues and opportunities surrounding neuronal diversity, in the context of neural diversity machines.

MODEL OF INTELLECTUAL VISUALIZATION OF GEOINFORMATION SERVICE

Stanislav L. Belyakov
Scientific and Technical Center
“Information Technologies”
Southern Federal University
347900, Taganrog, Russia
E-mail: beliacov@sfedu.ru

Alexander V.Bozhenyuk
Scientific and Technical Center
“Information Technologies”
Southern Federal University
347900, Taganrog, Russia
E-mail: avb002@yandex.ru

Marina L.Belykova
Scientific and Technical Center
“Information Technologies”
Southern Federal University
347900, Taganrog, Russia
E-mail: beliacov@yandex.ru

Igor N.Rozenberg
Public Corporation “Research and
Development Institute of Railway
Engineers”
Department of Computer Science
109029, Moscow, Russia
E-mail: I.rozenberg@gismps.ru

KEYWORDS

Visualization, geographic intellectual system, automatic mapping, uncertainty, fuzzy function, granulation.

ABSTRACT

In this paper we investigate a model of intellectual visualization of cartographic images. The selection of customer of geoinformation service most informative materials during the session is modeled. The use of a customer of geoinformation service of fuzzy function usefulness is a feature of the model. Model of selection of informative cartographical objects in the workspace of analysis is described. Estimation of level of usefulness uses the knowledge about the growth and reducing usefulness, witch depending on the number of objects in the cartographic image. The usefulness function is presented in a granular form. Cartographic description of the utility function is considered. Image defects due to mapping of partially defined situations are analyzed. These situations appear on the map due to the imperfections of algorithms of automatic recognition of real world objects. Visual and operational defects are marked. The model of the map visualization with defects of displaying of uncertain situations is built. The proposed approach will reduce the risk of making wrong decisions due to the incomplete and irrelevant maps of geographic information systems.

INTRODUCTION

Geoinformation services are a powerful tool for information support of decision-making processes. Management of objects and processes of the real world is based on the use of spatial data from electronic maps, diagrams and plans. Traditionally, geoinformation systems (GIS) are used for the storage and use of cartographic information. Geoinformation Services (GS)

is a geographic information system LAN or WAN. It continuously accumulates the map data and provides the access to this data through programming or interactive dialog interface. Google MAPS is a typical example of the geoinformation service.

Visualization is very important for map data when used in dialog mode. Visualization of map stimulates the intuitive creative thinking and promotes the use of deep knowledge of user-analyst. This mechanism is important for a decision making in difficult situations. A lot of complicated problems are solved in the process of visual studying of the maps (Egenhofer 2002, Erle et al. 2005, Cartwright et al. 2007).

Completeness, accuracy and relevance of map data are required for effective decision making. Low quality of maps generates the risk of damage due to an incorrect decision. However, the selection and study of useful data is problematic. The problem occurs for the following reasons:

- The map data redundant. None of application tasks needs all of the data simultaneously in the system. Analysts use data from local areas for solving the problems use;
- In a work session of analyst and GS the dynamics of change the operational set of data is present. Examining of images involves the addition and the removal of large volume of the map tiles;
- Analysts notion the quality of selected data is subjective.

As a consequence users spend a lot of efforts for management of the visual image. Quality of the basic problem solution is reduced inevitably. As any GS continuously accumulates the information about the outside world, this problem is escalating over the time. Because of the uncertainty and ambiguity of information visualization is of interest to the selection of useful information modeling of methods of artificial

intelligence. This question is little studied to date. Probably the reason is that the problem of selecting of useful map data is referred to the map construction (Li et al. 2013, Pettit 2008). It is assumed that any selected fragment of the map by the analyst automatically has maximum information. From our point of view, a selection mechanisms of a map data should be realized by GS. This will allow to newly solve the problem of timely updates for geographical maps. It is known that the period of their updates is standardly 5-7 years. In a number of applications this is totally unacceptable. An alternative option is to fill the GS database with «raw data» and use is in real time mode.(Konečný and Bandrova 2006).

This paper offers a model of visualization for useful map data selection in applied problem solving. Intellectual component of visualization process is defined by the accumulated knowledge (Luger 2004, Sowa 1999).

PROBLEM FORMULATION

The cartographic visualization task is seen as the process of map image of preset quality construction. This approach is different from the similar ones as it allows maximizing the efficiency of the cartographic images. It is expected that GS will generate the most efficient cartographic image. Herewith the generated output includes the uncoordinated objects and may be inconsistent with mapping standards. The survey shows that this kind of approach was never researched before.

The following tasks are to be solved when modeling the described process:

- To formalize the notion of cartographic image efficiency and to set objectives for visualization management;
- To fuzzy model the cartographic image efficiency;
- To model the visualization of uncoordinated and ill-defined objects for cartographic images.

MODEL OF VISUALISATION CONTROL

Solving the applied problem with the help of GS, the user implements a procedure of the cartographical analysis (Batty 2012, Dent 1990, Berlyant 1998). The creation of a local working area of global GS map is the basis of the procedure.

The working map area $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ that consists of map objects ω_i is a subset of objects $m_W \subseteq \Omega$ that describes the fragment of a map with boundaries

$$L_W = \{S_W, T_W, C_W, E_W\},$$

where S_W – spatial boundary, T_W – temporal boundary, C_W – semantic boundary, E_W – pragmatic boundary. The semantic boundary C_W is a set of classes of objects and relationships, which are described in the GS, E_W is a description of the limits of applicability of the map

constructed. The specific work area with boundaries is generated by the sequence of request from the client to the GIS server:

$$Q_i(X_S, X_T, X_C, X_E), i = \overline{0, N},$$

where X_* are, accordingly, the spatial, temporal, semantic and pragmatic parameters of a single request.

$Q_i(X_S, X_T, X_C, X_E)$ is a predicate of the request which in modern GS is built by user through the system of dialogue menus or is directly set as an expression in SQL.

The structure of the workspace $m_W \subseteq \Omega$ can be represented as the union of two sets:

$$m_W = B \cup E,$$

$$B \subseteq \Omega: \forall \omega_i \in B \Rightarrow \exists Q_j(X_S, X_T, X_C, X_E) = true,$$

$$j = \overline{1, N}, i = \overline{1, |\Omega|},$$

$$B \cap E = \emptyset, E \subseteq \Omega.$$

The set B is a skeleton of the request which is determined by predicates of requests, E – environment of the skeleton, in other words, it is subset of map objects that provides the continuity of maps (Berlyant 1997), and semantic entirety of the work area. The entirety is formed by using the expert rules of construction of the image of expert response $K(B, \Omega)$ to the skeleton of request B :

$$\omega_i \in E \Rightarrow K(B, \Omega) = true, i = \overline{1, |m_W|}.$$

Expert rules $K(B, \Omega)$ represent knowledge about how to construct the mapping images needed for solving the problem. Application of the rules $K(B, \Omega)$ leads to a reduction in redundancy of cartographic images and improvement of their informational content.

Informational content $I(m_W)$ of any work area is a related, subject determined value. Its value is validly estimated only in the narrow scope of a particular class of applied problems that can be solved by a certain group of users. General restrictions following from the specification of a person's perception of graphic images are as follows:

$$|m_W| = 0 \Rightarrow I(m_W) = 0,$$

$$|m_W| = \infty \Rightarrow I(m_W) = 0.$$

The level of information content cannot be estimated by any other way except for a set of expert rules $K_I(m_W)$. Rules reflect the subjectivity and ambiguity that are inherent to the evaluation of information content. Expert knowledge in the form of rules $K(B, \Omega)$ and $K_I(m_W)$ represent the "reasonable" rendering of map images strategy. The difference between the described intellectual visualization and the traditional

one is in support of subjective utility of map images. Not only the objects that meet the predicate of the request are visualized, but all those that fill a cartographic image with a meaning.

The management of visualization means solving the following problem:

$$\begin{cases} I(m_W) \rightarrow \max, \\ |m_W| \leq m^*, \\ m_W : Q_i(X_S, X_T, X_C, X_E) = \text{true}, i = \overline{1, N}. \end{cases}$$

m^* stands for the image complexity limit. The complexity is evaluated by the number of primitives.

To solve the problem, the GS should be equipped with software mechanism of expert evaluation $I(m_W)$ and a mechanism of adding and removing mapping objects in the working area. The GS works in a following way:

- The user authenticates himself at the beginning of a session with GS. This allows us to classify and choose it from a database of expert rules $K(B, \Omega)$, $K_I(m_W)$ and necessary thematic layers of a map. The effectiveness of further work in the session depends on how adequate classification is. The user might not be satisfied with the system performance. Thus the user may start a new session with other authentication parameters;
- The GIS server builds workspace $m_W = B \cup E$ with maximal information content for any client request. After that the server synchronizes the workspace with the client. Accounting of the limited resources of the client $|m_W| \leq m^*$ ensures the completeness of the result and security of interaction.

USABILITY ASSESSMENT APPROACH

Estimating the information content of the workspace $I(m_W)$ is the basis of the process of visualization management. In order to set the function of information content we have to take into account its dependency on the following factors:

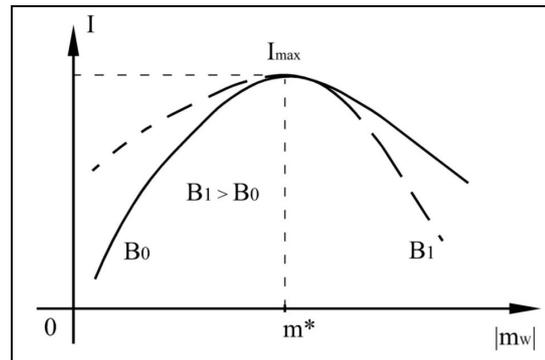
- the total number of map objects in the working area;
- the distribution of copies and classes of objects by the level of significance of the problem to be solved;
- the distribution by the level of importance of samples and classes of relations between objects;
- the degree of novelty in area studied for the user;
- the user's work experience with cartographic materials.

It can be said that information function depends on many variables $X = \{x_1, x_2, \dots, x_M\}$:

$$I(m_W) = F(X).$$

The number of variables is not known beforehand. The multidimensionality of the space of factors $\{x_1, x_2, \dots, x_M\}$ creates serious difficulties in finding the analytical relationship $I(m_W)$. Therefore it is proposed to evaluate the informative content by a fuzzy description. The description is based on expert data. This approach is explained by the high degree of incompleteness and uncertainty about the behavior of the function $F(x_1, x_2, \dots, x_M)$. The model of granular representation of the information function should be reviewed (Zadeh 1997).

The analysis of the analysts' behavior shows that the function $I(m_W)$ can be characterized by the curve in Fig. 1. The shape of the curve reproduces the qualitative features of the information content changes. With a small number of graphic objects in the workspace m_W information content is not high and grows as you add new objects. The growth rate of information content should be related to the adding of objects from the environment (E), because the information content of skeleton B represents the minimum level that the working space gets under the user's request. The dotted line shows the information curve for skeleton B_1 with a large number of objects ($|B_1| > |B_0|$).



Figures 1: The usefulness function

Where I_{\max} is the highest possible level of workspace's m_W information content. Qualitative analysis of the curve in Figure 1 allows us to enter at least three granules $\{I_L, I_M, I_R\}$. Each granule defines the area in space $I \times N$ in a fuzzy way:

$$\begin{aligned} I_L &= \{\mu_{Lk} / (i_{Lk}, N_{Lk})\}, I_L \subset I \times N, \\ I_M &= \{\mu_{Mk} / (i_{Mk}, N_{Mk})\}, I_M \subset I \times N, \\ I_R &= \{\mu_{Rk} / (i_{Rk}, N_{Rk})\}, I_R \subset I \times N. \end{aligned}$$

Where μ_{Ak} is the degree of belonging of the point (i_{Ak}, N_{Ak}) with the number k to a granule with an index A , an each pair (i_{Ak}, N_{Ak}) is a value of information content i_{Ak} with a number of objects N_{Ak} .

Please note that the definition of function in the space $I \times N$ is a simplification. Simplification is used deliberately to implement controls.

Granule I_L represents an area of increasing information content for cartographic images of low complexity of perception. Granule I_M is the area of the most informative images with the utmost level of perception. Granule I_R is the area of falling informative content of images difficult to perceive.

For the use and description of the granules $\{I_L, I_M, I_R\}$, we suggest using the cartographic representation. The essence of method is in constructing figurative and symbolic model of granules. The model describes the distribution of the importance of classes and samples of objects and relations of the work area; it also describes the relationship between the grade of information content and the total number of objects in the working space. Demonstrativeness of the cartographic representation, from our point of view, is extremely important in obtaining reliable and coordinated knowledge from expert GS users (Harrie and Weibel 2007).

Cartographic representation has a form of atlas – a set of maps that shows the behavior of informative value. Formally, this means that the information function is a superposition of functions

$$I(m_W) = F(x_1, x_2, \dots, x_M) = F(Y_1, Y_2, \dots, Y_H),$$

where

$$H < M, Y_i = Y_i(X_i), X_i \subset X, i = \overline{1, M}.$$

Each map of atlas shows the dependence $Y_i(X_i)$. The coordinate axes may be connected with different degrees of complexity. Each map is a graphic "projection" of the multidimensional space of problem domain concepts. The problem of mapping $I(m_W)$ thus is reduced to the determination of a set of "projections" with the required properties.

Let us consider the example of mapping. We assume that the atlas consists of two maps: the first allows determining a preliminary assessment of informative value, the second determines the granule, which contains the analyzed workspace of the map. Technically

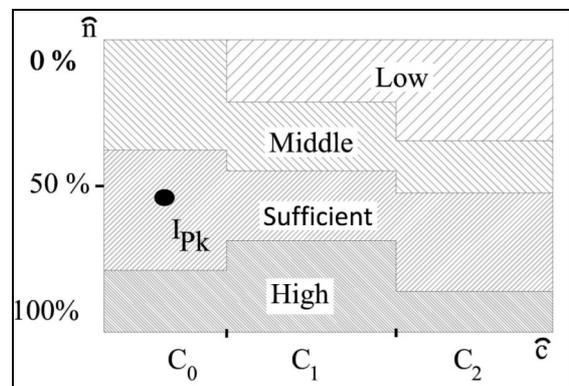
$$I(m_W) = F(I_P, |m_W|), I_P = Y_1(\hat{n}, \hat{c}).$$

The value I_P is determined from the set $\{\llcorner Low \llcorner, \llcorner Middle \llcorner, \llcorner Sufficient \llcorner, \llcorner High \llcorner\}$. Any value of I_P subjectively displays a preliminary evaluation of informative value. The value of I_P depends on the use of layers, objects and relations of the workspace map. Fig. 2 shows the example of a map that binds used layers, objects and relationships that are clustered into

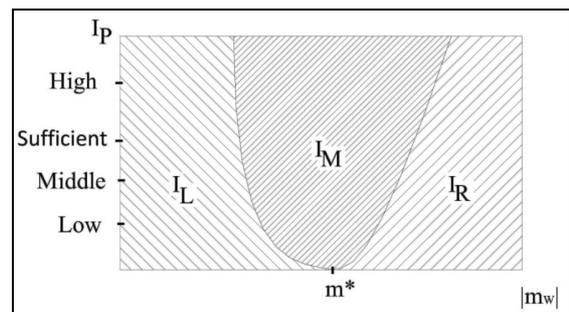
three clusters $\{C_0, C_1, C_2\}$. The coordinates of any point on the map is (\hat{n}, \hat{c}) where \hat{n} is a fuzzy evaluation of the relative number of objects and workspace relations, \hat{c} is the name of the most significant cluster of objects and relations in the work area. For example in Figure 3 point I_{Pk} displays workspace that contains some $\hat{n} = 40\%$ elements of the cluster $\hat{c} = C_0$. The cluster C_0 includes:

- examples of "automobile filling station";
- examples of the relation "nearby service stations";
- layers of "parking lots", "motorway", "restaurants", "service station", "hotels", "ATM";
- relations "crosses", "attached", "located in the danger zone."

The map of preset informative content is zoned. Each area corresponds to one of the values of the set $\{\llcorner Low \llcorner, \llcorner Middle \llcorner, \llcorner Sufficient \llcorner, \llcorner High \llcorner\}$. Figure 3 shows a map of zoned granules of information content $F(I_P, |m_W|)$.



Figures 2: The map of preliminary informativeness



Figures 3: The map of the arrangement of information content granules

In this example of the map the fuzziness of granules is not displayed for the sake of simplicity. The coordinates of any point $(I_P, |m_W|)$ clearly belong to a single granule. Having determined I_{Pk} on the map (Fig. 2), counting the total number of objects in the workspace m_W , you can accurately match the workspace to the granule of information content on the map (Fig. 3). All procedures are implemented in software.

VISUALIZATION FOR PARTIALLY DEFINED SITUATIONS

To get the operational spatial data one needs to access heterogeneous systems that record things and events of the real world. The Internet plays a special role in that situation. For example, the information about the natural anomalous situation can be got as a message from the news flow (RSS), published space or aerial photos, video streams from Web-cameras, reports from the electronic media, from personal blogs, from specialized communities in social networks, as well as from the map services. Quite often this information is not metric and does not contain an explicit gridding. However, its value in the case of responsible decision-making is very high. The information values may compensate the lack of accuracy and completeness.

Let us assume that GS has a search engine that can find information resources for later retrieval of spatial data. GS also has a set of programs for identification and mapping of situations. It is well known (Harrie and Weibel 2007) automatic mapping is not perfect. Newly created mapping objects can disrupt the logic of the map. Let us consider the model of visualization for the case when the objects inconsistent with the other map elements are added to the map.

Let us describe the set of mapping objects built by the recognition of situations outside the world as $S = \{s_i\}$. The information base of GIS is complemented by multiple map objects S in the process of real-time mapping:

$$\bar{\Omega} = \Omega \cup S.$$

Unlike any other object $\omega_i \in \Omega$, the space-time coordinates and semantic attributes $s_i \in S$ are not completely reliable because of the limited capacity of the recognizing subsystem.

Adding context to the working area and its further visualization with maximal information content

$$I(B \cup E \cup S)$$

cannot be performed by the method discussed above. The reason is that the visualization of situations $s_i \in S$ creates defects of cartographic displaying. The defect appears, for example, like an overlay of images of situations $V(s_i)$ and objects of cartographic basis:

$$V(s_i) \cap V(\omega_j) \neq \emptyset.$$

Analysis has shown that defects in the mapping image can be divided into two classes.

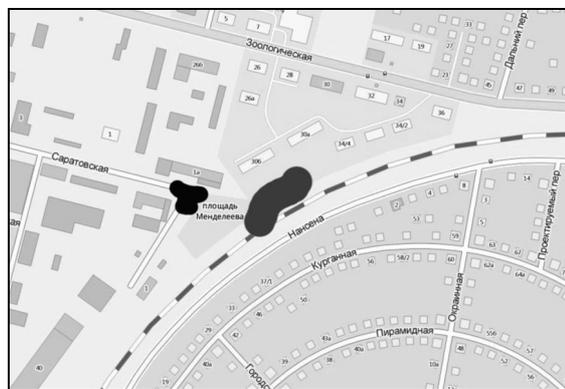
Defects of the first class are visual defects that create a misconception about the objects with overlapping images. But in this case any of the overlapping objects of the map would be

perceived incorrectly under the visual analysis. Fig. 4 shows an example of a map with the representation of two situations: the both are displayed as areal objects of different colors. The situation depicted in gray overlapped with the line of the railway. However, the map is correctly recognized by the analyst. Contrary to that, the situation displayed by an area objects in black, hides an important object to analyze – the bend of the road. The visual recognition of the road would be incorrect.

Operational defects form a second class. Defects of this class produce incorrect results when the procedures of cartographic analysis are performed. The example in Figure 4 shows how an error of estimation of the topological relation "to be around" occurs. If you request "to find all objects that are close to the railway", the result will be incorrect. The object of situation shown in gray is for "overlap". The performance of a particular GS determines the possibility of operational defects. It may happen that the GIS has no evaluation procedures of topological relations. Then the operational defects are not possible.

Despite the fundamental differences in the causes of occurring and influence of defects, we can claim that any defect reduces the productivity of the cartographic image. In this case the procedure of productivity maximization can make one of two actions:

- remove the object $s_i \in S$ that raises the defect;
- correct the defect by removing the objects $\omega_i \in \Omega$.



Figures 4: The example of image with situation $s_i \in S$

Let us estimate the relevance of each action. Suppose that after the elimination of defects the set $D = \{d_i\}$ of objects removed from the workspace is formed. Then the modified working area described by the set

$$\tilde{m}_W = B \cup E \cup S \setminus D.$$

Adding indefinite situations to the workspace has the effect, if the following inequality is right:

$$I(B \cup E \cup S \setminus D) > I(B \cup E).$$

Finally, the procedure of visualization control can be described as follows:

- to build for the next request $Q_i(X_S, X_T, X_C, X_E)$ skeleton B and the environment E ;
- to assess the information content of the constructed workspace;
- to construct objects $S = \{s_i\}$ that display information from external sources;
- to identify the visual and operational defects and to find objects that can be removed and thus that removal will increase the information content of cartographic images by a linear search;
- to synchronize the workspace with the GS client.

CONCLUSION

The approach described above results in the following. The main factor for smart visualization quality improvement is the adequate balance between the expert description and the subjective perception of cartographic images. The productivity function when displayed in granules represents the unclearness and incompleteness inherent to the expert knowledge on productivity on the whole.

An important factor in the quality of smart visualization is the presence of a GS engine that allows extracting spatial data from data sources of different nature. The more important the data is the greater deviation from the standards of cartographic representation comes. The proposed approach puts the visual analysis of maps in the first place. A powerful mechanism inherent to men – an intuitive understanding of the meaning of the map is activated with the support of GS.

This approach gives the following positive results.

We assume that the wrong decision leads to damage U . The damage occurs when the map contains no objects $S = \{s_i\}$. The damage is reduced if such objects appear on the map and analyst successfully identifies the state of the outside world. The information content is maximized and it can be assumed that an absolutely accurate identification is determined by the effectiveness of recognition programs. We denote p by the probability of correct recognition of a given situation. Then, using the formula of the binomial distribution

$$P_N(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

we can calculate the probability of that k situation would be recognized from N actually existing ones. Mean cost in this case is calculated as

$$\bar{U} = (1 - P_N(k))U.$$

This expression allows you to evaluate the effect depending on the parameters of the recognition system (probability p), and the specific of the decision making process (values N and k).

Acknowledgements. This work was partially supported by the Russian Foundation for Basic Research, projects N 13-07-13103 ofi_m_RzhD, and N 12-07-00032.

REFERENCES

- Batty, M. 2012 “A Generic Framework for Computational Spatial Modelling”. In: *Agent-Based Models of Geographical Systems*. Springer Science+Business Media B.V.
- Berlyant, A.M.1998.” ’Not For Use To Foresee...’, Or About Future Maps”. *Mapping Sciences and Remote Sensing*. V.H. Winston & Sons, Inc. No.3(35),166-172.
- Cartwright, W.; M.P Peterson; and G. Gartner. 2007. “Multimedia cartography”. Springer Verlag, Berlin.
- Dent, B.D. 1990. “Cartography: Thematic map design, 2nd edn”. Wm. C. Brown Publishers, Dubuque, Indiana.
- Egenhofer, M.J. 2002. “Toward the Semantic Geospatial Web”. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS-2002)*. Virginia: ACM Press, 1–4.
- Erle, S.; R. Gibson; and Walsh, J. 2005. “Mapping Hacks – Tips & Tools for Electronic Cartography”. Sebastopol, CA, O’Reilly Press.
- Harrie, L. and R. Weibel. 2007. “Modelling the overall process of generalization”. In: *Ruas.A., Mackaness, W.A. and Sarjakoski, L.T. (eds.): Generalization of Geographic Information: Cartographic Modelling and Applications, Series of International Cartographic Association*. Elsevier, 67–87.
- Konecný, M. and T. Bandrova. 2006. “Proposal for a Standard in Cartographic Visualization of Natural Risks and Disasters”. In *Proceedings of the Joint Symposium of Seoul Metropolitan Fora and Second International Workshop on Ubiquitous, Pervasive and Internet Mapping*. Seoul, Korea, 165–173.
- Li, W.; L. Li; M.F. Goodchild; and L. Anselin. 2013.”A geospatial cyberinfrastructure for urban economic analysis and spatial decision-making”. *ISPRS International Journal of Geo-Information*, No.2, 413–431.
- Luger, G.F. 2004. “Artificial Intelligence: Structures and Strategies for Complex Problem Solving”. Addison Wesley.
- Pettit, C.; W. Cartwright; I. Bishop; K. Lowell; D. Puller and D. Duncan.2008. “Landscape Analysis and Visualisation, Spatial Models for Natural Resource Management and Planning”. Springer-Verlag, Berlin.
- Sowa, J.F. 1999. “Knowledge representation: logical, philosophical, and computational foundations”. Brooks Cole Publishing Co, Pacific Grove.
- Sui, D.; S. Elwood; and M. Goodchild. 2012. “Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice”. Springer.
- Zadeh L.1997. “Towards a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic”. *Fuzzy Sets and Systems*, No.90, 111-127.

AUTHOR BIOGRAPHIES



STANISLAV L. BELYAKOV was born in Ukraine. He graduated Leningrad Electrotechnical Institute in 1982. Professor of Scientific and Technical Center “Information Technologies”, South Federal University, Russia. He holds a degree of Doctor of Technical Sciences in Theoretical Foundations of Informatics in 2003. His research interests include geographic information systems, modeling intelligence. Contact information: 44, Nekrasovsky Street, Taganrog, 347900, Russia, phone: +78634371743, e-mail: beliacov@yandex.ru.



ALEXANDER V. BOZHENYUK was born in Taganrog, Russia. Professor of Scientific and Technical Center “Information Technologies”, South Federal University, Russia. He holds a degree of Doctor of Technical Sciences in Theoretical Foundations of Informatics. His research interests include fuzzy models, fuzzy decision making. He has more than 180 publications in the fields of Computer Sciences. Contact information: 44, Nekrasovsky Street, Taganrog, 347900, Russia, phone: +78634371743, e-mail: avb002@yandex.ru.



IGOR N. ROZENBERG was born in Taganrog, Russia and went to the Radio Engineering Institute. He holds a degree of Doctor of Technical Sciences in Geographic Information Systems in 2007. He is the First Deputy General Director of Public Corporation “Research and Development Institute of Railway Engineers”. Contact information: 27/1, Nizhegorodskaya Street, Moscow, 109029, Russia, phone: +78634371743, e-mail: I.rozenberg@gismps.ru.



MARINA L. BELYAKOVA was born in Taganrog, Russia. She graduated Leningrad Electrotechnical Institute in 1984. Received a Ph.D. in 1989 in Taganrog Radio Engineering Institute. Contact information: 44, Nekrasovsky Street, Taganrog, 347900, Russia, phone: +78634371743, e-mail: beliacov@yandex.ru.

ADVANTAGES OF USING MEMETIC ALGORITHMS IN THE N-PERSON ITERATED PRISONER'S DILEMMA GAME

Tamara Álvarez, Miguel Loureiro, José Covelo, Ana Peleteiro, Aleksander Byrski, Juan C. Burguillo
Department of Telematics. E.E. de Telecomunicación
Universidade de Vigo
36310-Vigo. SPAIN
E-mail: J.C.Burguillo@uvigo.es

KEYWORDS

Agent Based Simulation, Simulation of Intelligent Systems

ABSTRACT

Memetic algorithms are a type of genetic algorithms very valuable in optimization problems. They are based on the concept of “meme”, and use local search techniques, which allow them to avoid premature convergence to suboptimal solutions. Among these algorithms we can consider Lamarckian and Baldwinian models, depending on whether they modify (the former) or not (the latter) the agent's genotype. In this paper we analyze the application of memetic algorithms to the N-Person Iterated Prisoner's Dilemma (NIPD). NIPD is an interesting game that has proved to be very useful to explore the emergence of cooperation in multi-player scenarios. The main contributions of this paper are related to setting the ground to understand the implications of the memetic model and the related parameters. We investigate to which extent these decisions determine the level of cooperation obtained as well as the memory and the execution performance.

1. INTRODUCTION

Memetic algorithms emerge at the end of the 80s inspired both in the Darwinian natural evolution principles, and in the “meme” concept introduced by Dawkins (Dawkins 1976). Dawkins defines a “meme” as an information unit that reproduces with people's exchange of ideas. This term is the cultural evolution equivalent of the gene concept, and in the case of memetic algorithms case, it refers to the strategies used to improve individuals' knowledge or culture.

Memetic algorithms are also called hybrid genetic algorithms, or genetic local searchers, and can be defined as a type of genetic algorithms, which introduces local search techniques carried out by each agent of the population individually. Thus, memetic algorithms combine exploration with exploitation abilities provided by the local search, and they reduce the premature convergence to local optima due to a better exploration of the solutions space. Thus, they are very efficient in optimization tasks, like multi-objective optimization (Knowles and Corne 2000), combinatory optimization (Garg 2009), as well as other applications (Krasnogor and Smith 2003).

The study of cooperation evolution in the Prisoner's Dilemma has been extensively studied by means of genetic algorithms, and classic strategies by many authors like Axelrod (Axelrod 1984) or Xin Yao (Yao and Darwen 1994). The latter article has been used as a reference for the development of the experiments introduced in this paper.

The main contribution in this paper is to explore the possibilities of memetic algorithms for simulating the NIPD. We investigate different scenarios for the game, and explore the possible combinations of algorithms and game parameters, that produce better outcomes from the cooperation point of view.

The rest of the paper is organized as follows. Section 2 presents the some background necessary to understand this work. Section 3 presents a description of the memetic model used in the NIPD evolutionary game. Section 4 presents the results that allow us to analyze the performance of some memetic algorithms and the game conditions. Section 5 draws the conclusions obtained, and hints some possible future work we foresee from this point.

2. BACKGROUND

In this section we introduce several topics needed to understand the NIPD game, the model and the scenarios described in this paper.

N-Iterated Prisoner's Dilemma (NIPD)

The Iterated Prisoner's Dilemma (IPD) consists on repeating successively a set of basic Prisoner's Dilemma games (Axelrod 1984). Players take their decisions in each round considering a certain number of previous opponent' actions without knowing the number of rounds to be played. This prevents individuals from building their strategies depending on a certain time horizon. A strategy in the IPD is a rule to decide the next action depending on the previous history. The success of a strategy depends not only on that strategy, but also on the opponent's strategies.

The basic 2-player IPD has been used to model several real-world problems. However, there are other scenarios that cannot be modeled with the 2-player IPD (2IPD). The N-player IPD (NIPD) is used to model situations where one player interacts with more than on opponent at the same time. In the NIPD game, each player can also choose between Cooperation (C) and Defection

(D), but the selected strategy is used combined with the ones selected by the rest of the opponent2s.

A possible payoff matrix for a N-Player IPD is shown in Table 1, where in each iteration, each player receives a payoff that depends on its action and how many individuals have cooperated in a particular game iteration.

Table 1: NIPD Payoff Matrix
Number of Cooperators

		0	1	2	...	$n - 1$
Player's action	C	0	2	4	...	$2(n - 1)$
	D	1	3	5	...	$2(n - 1) + 1$

Genetic Model

An individual's strategy determines which action will be performed in a certain situation. When using genetics, we can represent the strategy using a genome, i.e., an array of bits that specifies the decision to take in every possible context. We will base our work in the genotypic approach introduced by Xin Yao in (Yao and Darwen 1994).

As Xin Yao et al. explain in their model, each individual is regarded as a set of rules stored in a look-up table that covers every possible history. As a game has an enormous number of possible histories, and as only the most recent steps will have significance for the next move, we only consider every possible history over the most recent h steps, where h is less than 4. This means that an individual can only remember the h most recent rounds. Such a history of h rounds is represented by:

1. h bits for the player's own previous h moves, where a "1" indicates defection, "0" cooperation
2. Another $n-1$ group of $h \cdot \log_2(n)$ bits for the number of cooperators among the other $n-1$ players, where n is the number of the players in the game. This requires that n is a power of 2.

For example, if we are looking at 8 players who can remember the 3 most recent rounds, then one of the players would see the history as: History for 8 players, 3 steps: 001 111 110 101 (12 bits) were the first 3 bits on the left represent player's own actions (see (Yao and Darwen 1994) for a detailed example). In this example we have $2^{12} = 2048$ possible histories, so the same number of bits are needed to represent all possible strategies. In the general case of an n -player game with history length h , each history needs $h + h \cdot \log_2(n)$ bits to represent and there are $2^{h + h \cdot \log_2(n)}$ of such histories. Since there are no previous h rounds at the beginning of a game, we need to specify them with another $h \cdot (1 + \log_2(n))$ bits. Hence each strategy is finally represented by a binary string of length $h + h \cdot \log_2(n) + h \cdot (1 + \log_2(n))$.

The different strategies used by the players, and represented by this model, form the genetic diversity of the experiment. Introducing other factors, like

mutations, increases this genetic diversity. The measurement of this diversity can be very important to analyze genetic algorithms, and can be used as a finishing condition when the genetic diversity falls below a certain threshold. There are many methods to measure this diversity, e.g., the entropy, the Hamming distance, and the moment of inertia. In (Morrison and De Jong 2002) Morrison and De Jong showed that the moment of inertia obtains the same results to measure diversity than the Hamming distance, but with a much lower computational cost, and it is the approach that we use in this paper.

Memetics: Lamarckian and Baldwinian Models

Memetics follow a process similar to genetic algorithms. Both keep a population of different solutions to the problem considered, and perform processes of evaluation, selection, crossover and mutation. However, memetics incorporate the novelty of a local search performed by each individual. This new feature implies that individuals can find individuals fitter than themselves who are close to them in the search space.

The Lamarckian model (Nghia et al. 2009) is based on the fact that the improvements found in the local search process are assimilated by the individual, transmitting the positive characteristics directly to the offspring. This means that the fittest neighbor found substitutes the original candidate. This model is applied along with the traditional mutation operator after the processes of evaluation; selection and crossover have been made.

On the contrary, the Baldwinian model (Castillo et al. 2006) only changes the individual's fitness, and the improved genotype does not become part of the population. This is, learnt behaviors are not transmitted directly to the offspring. Instead they are transmitted in an indirect way incorporating them in the fitness value. Therefore, it can be seen as a search for those individuals that could produce the fittest children, this means the best potential solutions to the problem. This process is carried out generating through mutation a predefined number of descendants. Each one of them competes against the $N-1$ rivals and the best fitness obtained is assigned to the parent. This process represents learning and improvement of an individual.

Spatial and Panmitic Populations

The populations considered in this paper are organized in panmitic or spatial networks.

In panmitic scenarios each individual can interact with any individual of the population. Spatial networks (Knowles 2000) are a collection of nodes connected by links, with a certain relation determined by the spatial proximity among them. Each node, representing a certain player, is connected to a group of other nodes, i.e., neighbors that form its neighborhood. Therefore, neighborhood is set by a radius that defines the maximum distance a neighbor can be. This configuration represents relationships conditioned by

proximity, and it is related with regular graphs. The use of a spatial organization of the population has been studied in different studies of the Prisoner Dilemma using 2 players (Nowak and May 1993), using the classic game (Schweitzer et al. 1997), using classic strategies (Nakamaru et al. 1997) or analyzing factors like the influence of the payoff matrix or the cooperation evolution. An example of these networks can be seen in Figure 1.

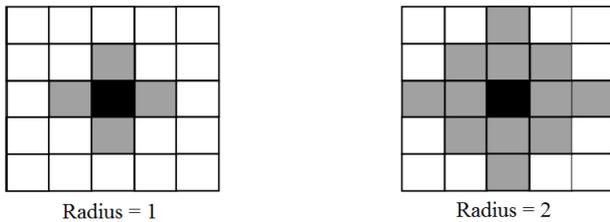


Figure 1: Neighborhood Representations (in gray)

The Island Model

Despite how helpful evolutionary algorithms can be, there are still problems that require a great amount of execution time. Parallel programming techniques have been developed to tackle this limitation. Parallel evolutionary algorithms can run several genetic algorithms at the same time using small populations, solving this way a problem of higher complexity and reducing the time to wait for a result.

The Island Model is a type of parallel evolutionary algorithm consisting of dividing a large population into smaller subpopulations or islands. Each of these islands is an initial population of an evolutionary algorithm, which is executed in parallel. In this model each of these islands have the possibility of exchanging the best individuals (distributed system), or not exchanging anything (partitioned system).

The movement between different islands receives the name of migration. Migrations exploit the differences between the different subpopulations, since as all of them start with different initial populations, convergence will take different paths in each of them.

The number of migrations is a point to consider, since a migration of a large number of individuals implies a reduction in the global diversity, and therefore removing the differences among islands. On the contrary, when the migration rate is too low it could lead to an early convergence of the subpopulations.

3. MODEL DESCRIPTION

The algorithm used in our simulations starts with an initialization process where an initial population is created. After this initialization, the evaluation and breeding take place until diversity is too low or a maximum number of iterations is reached. After the evaluation and breeding processes there is an optional phase of exchange, used in the Island model, where individuals are exchanged between islands.

The agents will play NIPD games using one of the memetic models described in Section 2. Each generation plays several rounds selecting N random players. If we are in a Baldwinian model we generate a certain number of descendants to compete against the rivals, and we assign the best fitness obtained to the parent without modifying the genotype. In case of a Lamarckian model in this point the game is played naturally between the agent and its opponents.

Once all the individuals are evaluated, and have their fitness calculated according to the Payoff matrix described in Table 1, the breeding process starts. This stage consists of a process of selection, crossover and mutation.

To select the breeding parents we use the roulette selection. This method selects N individuals randomly (one individual may be selected more than once). After this selection only the individual with the best fitness will be chosen (or the worst fitness if configured with that criterion).

The Island model is introduced in three different places. Before the selection and modification stage there is a new process that extracts individuals from each Island and stores them in a remote process. After the breeding stage other process is in charge of introducing the extracted individuals in the corresponding islands. Lastly, there is a last process in charge of checking if the simulation must be stopped because one of the parallel islands has found the individual wanted. The Island Model used in this paper considers each island a subpopulation of the initial population and all the subpopulations must reach an end separately.

The methods used to select the individuals sent to other islands, and the individuals removed when others arrive from other islands, are Tournament and Random selection respectively.

4. RESULTS

This section presents and analyzes the results obtained in the experiments performed with our model.

Experimental Settings

The parameters used in the simulations are indicated in Table 2, unless otherwise indicated. When there is no spatial distribution considered, the population distribution will be by defect panmictic, i.e., each individual can interact with any other.

Table 2: Default parameters used in the experiments

Parameter	Value
Number of Simulations	10
Number of Generations	1000
Number of Subpopulations	1
Size of the Subpopulations	100
Size of Tournament	2
Mutation Probability	$1/L$ (L = genotype size)
Crossover Type	Uniform (0.5 probability)
Num. of plays in Memory	$\{1, \dots, 5\}$
Spatial Distribution	$\{\text{yes, no (panmictic)}\}$

Lamarck vs. Baldwin Comparative

In this section we will study the differences between the Baldwinian and the Lamarckian models, analyzing the influence of parameters related to the memetic local search process.

Figure 2 offers a representative result of our comparatives among Baldwinian, Lamarckian and genetic payoffs for different number of players and history sizes.

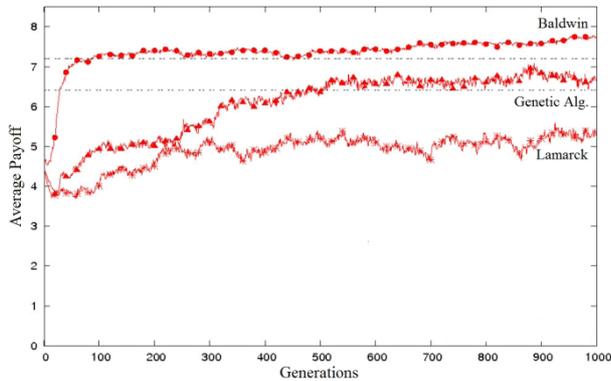


Figure 2: Average Payoff in a 5IPD with 2 plays History

It can be seen that the Baldwinian algorithm obtains a level of cooperation over 90%. However, Lamarckian and genetic algorithms can only reach much lower levels. In the Lamarckian case this is due to the fact that changes in the population are introduced continuously with new and random individuals, which implies an increase in the population diversity. Thus, while in the Baldwinian model population diversity decreases more and more until in the end, when cooperative strategies dominate; in the Lamarckian model it cannot be reduced to the same extent. This diversity reduction can be seen more clearly in Figure 3.

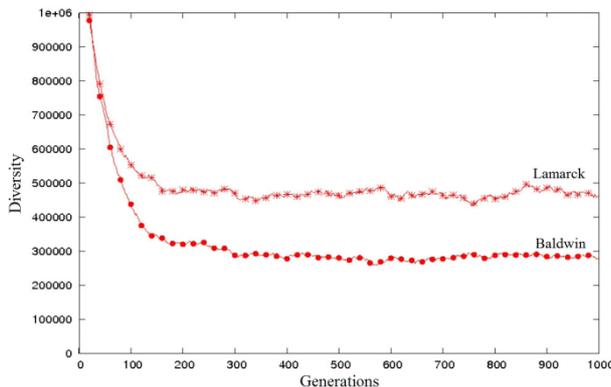


Figure 3: Avg. Diversity in 4IPD with 3 plays History

In our simulations we have found that the Baldwinian and the Lamarckian models seem to be more useful than simple genetic algorithms in cooperation evolution. They show a better performance in some cases due to the memetic local search.

Considering the results summarized by these figures, the following experiments will make use only of the Baldwinian model, since the Lamarckian one introduces

a higher level of population diversity, which could be an advantage in other studies (Nghia et al. 2009), but that does not lead to a better performance in the case of evolutionary algorithms in dynamic environments like the ones used in this paper.

Influence of the Number of Players

The first parameter that requires analysis is the influence of the number of players that participate in the NIPD game. To study this parameter we have used a Baldwinian model with a 2 plays History where we analyze how the cooperation level changes when the number of players varies.

In the next figures we have represented in the x axis the number of generations and average payoff in each generation in the y axis. Since the graphs are associated to different y axis values, we have also included two reference level marks so that they can be more easily compared. These levels refer to the 80% and 90% of the maximum value that can be obtained, when all the individuals cooperate.

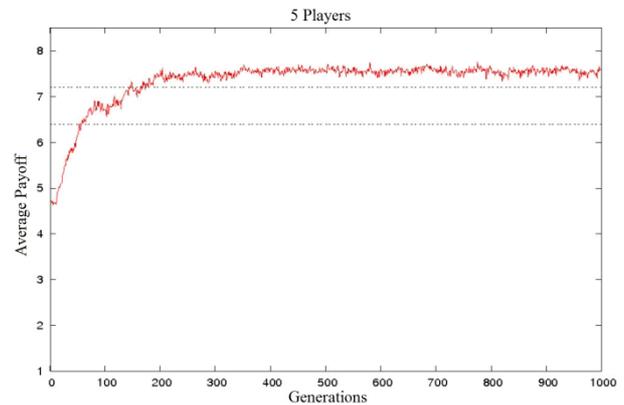


Figure 4: Average Payoff for 5IPD and 2 plays History

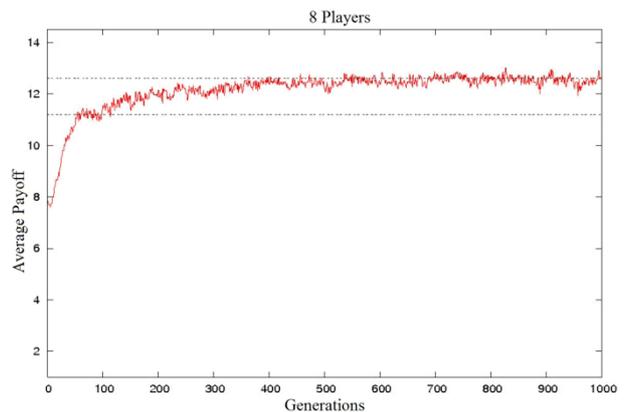


Figure 5: Average Payoff for 8IPD and 2 plays History

In Figures 4 and 5 it can be seen that as the number of players increase, reaching a certain level of cooperation becomes increasingly harder.

The 80% of cooperation is reached rapidly in both cases, but the as the group becomes larger it becomes more difficult to reach a level over 90%.

The average number of cooperators and defectors per round and generation is therefore affected by the number of players involved. This result matches the conclusions obtained by Xin Yao in (Yao and Darwen 1994).

Influence of the Number of Mutations

Each time an individual is evaluated, it is mutated x times. Each one of those mutations is mutated again another x times until a parameterized number of generations. In Figure 6 we present a comparative of the payoff obtained with different number of mutations and generations.

The results show that the 8 mutations and 2 generations case obtains better results. Despite reaching the 90% cooperation level in all cases, the highest values are obtained faster when the number of mutations evaluated is higher. This is explained by the fact that in the case with more mutations and generations, the local search takes into account more possibilities.

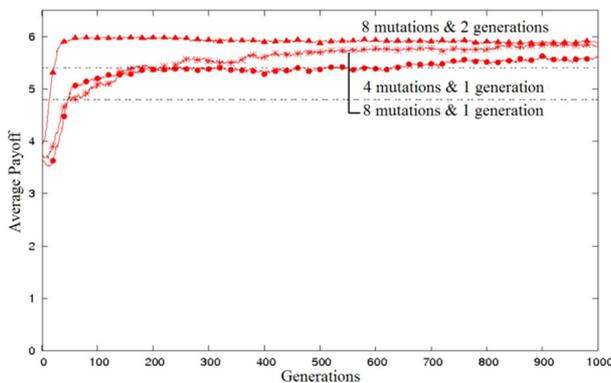


Figure 6: Average payoff in 4IPD with 3 Plays History

Influence of the History Size

As the number of actions remembered and the number of players increases, the individual's genotype becomes bigger according to the explanation in Section 2. Thus finding a suitable value is important since it affects the execution time and the memory used.

In Figure 7 we can see that with a memory of only 1 play the results are quite satisfactory with a cooperation level between 80 and 90%. The best results are obtained with 2 and 3 plays, but when the memory size is increased over 3 plays the performance is becoming to be impaired. These results are explained by the fact that with one play the information is insufficient to decide the most convenient action to take. With 2 or 3 plays an individual can act taking into account more actions and avoiding misjudging the opponent. However, when the memory size is greater than 3 individuals can take into account defections occurred a long time before the current play. Besides, as the number of combinations becomes higher, the genotype size becomes larger too. This means that the population diversity is increased as well. These two factors contribute to impair the cooperation as the history size increases.

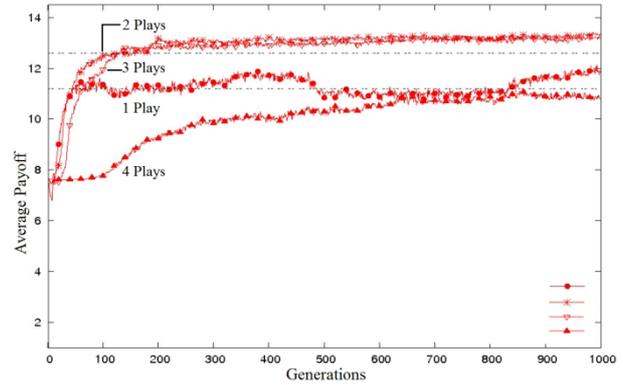


Figure 7: 4IPD Avg. Payoff with different History sizes

Spatial Networks: Players Selection

The previous experiments have considered only panmictic populations. Figure 8 shows the influence of using a spatial organization to select the users' opponents instead of allowing connections with any user of the population.

As it can be seen in figure 8, in the case of a population organized in neighborhoods cooperation appears faster, reaching the 90% level many generations before than in the panmictic case.

Figure 8 shows a superposition of 10 simulations with 8 players and 2 plays history for random (left) and spatial distributions (right).

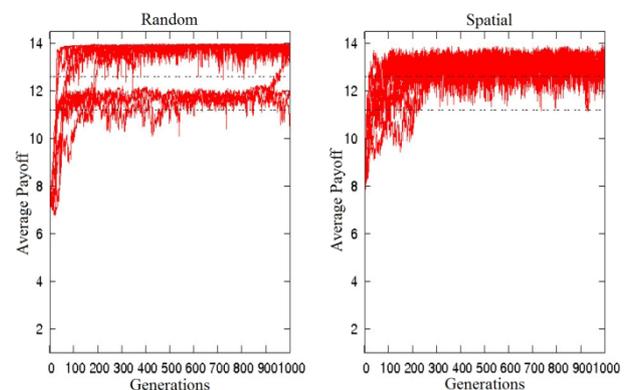


Figure 8: 8IPD with 2 plays history for Random (left) and Spatial (right) Distributions

These results show that after 1000 generations, in the random distribution case not all the simulations reach the 90% while in the spatial case all of them do it.

Despite organized networks render better results, as it has just been seen, there are certain experiments in which the cooperation is not increased. This happens in those cases where a high level of cooperation is already reached with a panmictic distribution and therefore it becomes difficult to improve even introducing a spatial organization (e.g. the example shown in the Influence of History Size subsection).

We have selected individuals in spatial networks using methods like best neighbor selection, selection by tournament between the individual's neighbors and substitution by the best neighbor after an evaluation.

Figure 9 shows a comparative of these methods in a 4IPD with 1 play memory size game.

The results show that both the best neighbor selection and the substitution by the best neighbor perform well, while the tournament selection is the one that shows the worst performance.

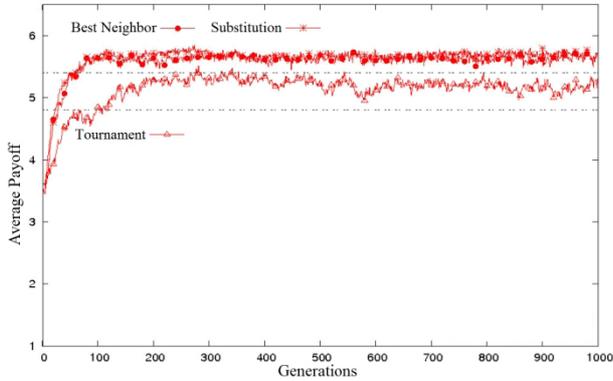


Figure 9: Selection Methods in 4IPD with 1 play History in Spatial Networks

Spatial Networks: Neighborhood Size

An important subject in Spatial Networks is the neighborhood size. Figure 10 shows how this parameter affects the performance of a 5IPD with a history size 2 and 3 respectively in a 10x10 spatial population.

Analyzing these graphs it is possible to conclude that the smaller the neighborhood, the easier the cooperation can thrive. This result also agrees with the conclusions of Xin Yao (Yao and Darwen 1994).

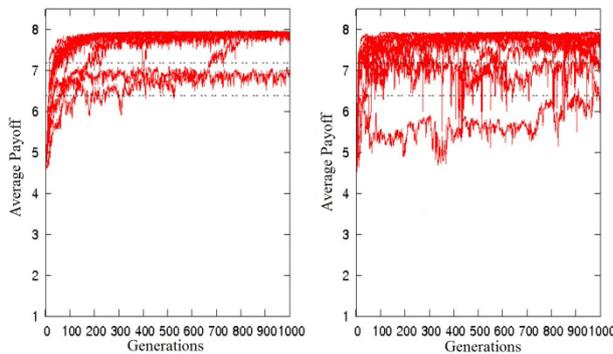


Figure 10: 5IPD with 3 plays History in Neighborhoods of Size 4 (left) and 30 (right)

The Island Model

As we have seen in Section 2, the Island Model helps solving problems where the time of execution becomes a serious constraint.

In this model the initial population is divided in subpopulations, where each one of them will evolve independently while every certain period of time they exchange individuals (migration).

Since the subpopulation initialization is random different results can be obtained in each subpopulation leading to differences in the evolution of cooperation among individuals. Besides, when the migration rate is

too low, or there is none, islands work as independent simulation runs.

The following experiments have been done with groups of 4 players and a memory size of 3 plays with spatial distribution and using the best neighbor selection. The population will be organized in 4 subpopulations with identical characteristics and 100 individuals in each one of them. The parameters used are shown in Table 3.

Table 3: Parameters for the Island Model Simulations

Parameter	Value
Topology	Ring
Number of Individuals Exchanged	5 (5% of the population)
Migration initial Generation	20
Period between Generations	20
Number of generations	1000
Number of simulations	10

Parameters like the number of islands a subpopulation sends individuals to and which are them, are determined by the topology of the model (Jovanovic et al. 2010)(Rucinski et al. 2010). Other parameters that we will change are the number of individuals exchanged, the generation to start exchanging, and the period between migrations.

The moment when the migrations begin is important, since if it is too early the search is not aimed correctly yet and therefore the individuals exchanged will not be fit for the problem to evaluate.

The effect of the migration rate and the topology on the cooperation can be seen in the following figures.

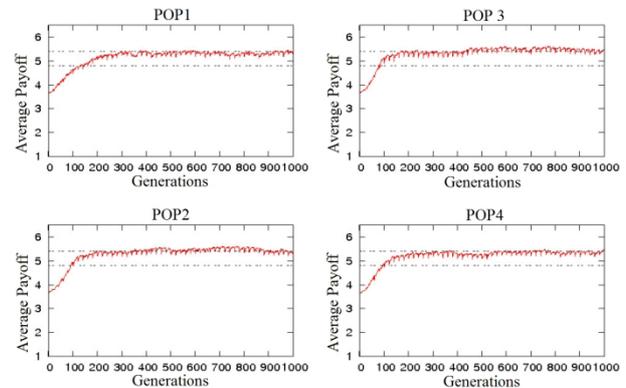


Figure 11: Ring Topology and 5% Migrations

In Figure 11 we can see that, as the migration rate is so low, each subpopulation evolves in the same way as if they were a unique independent population. Thus, in this case the migrations do not have a significant contribution to the evolution.

Figure 12 shows the same simulation increasing the migration rate to 50% of the population (50 individuals). When the migration rate is so high the cooperation evolution is considerably affected, since every time that a subpopulation finds fit individuals they are removed from the island in a migration.

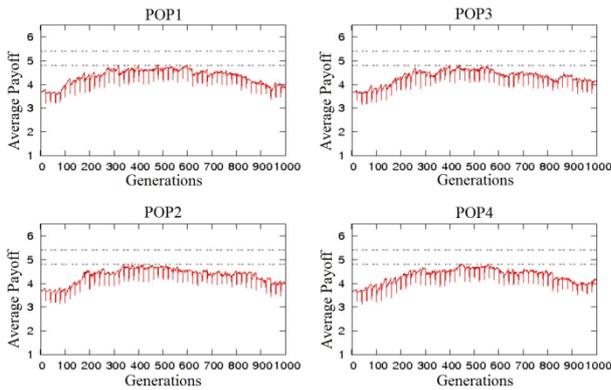


Figure 12: Ring Topology with 50% Migrations

Finally, in figure 13 we consider a star topology where the central island will be POP1 with a migration rate of 90%, while the other islands will have a 10% migration rate. In this case we can see that the central island does not reach high levels of cooperation while the other islands achieve good levels. This behavior is due to the fact that the migration rate is too high and we are introducing new individuals constantly in the central subpopulation.

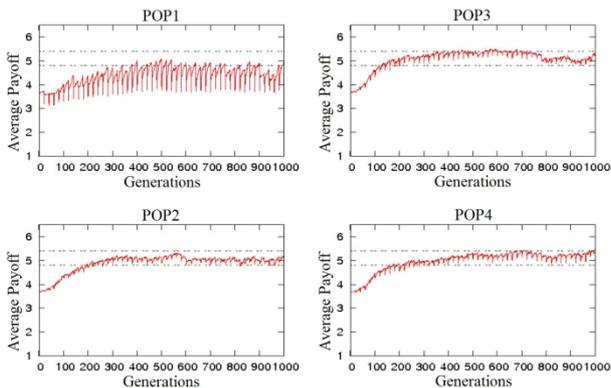


Figure 13: Star Topology. 90% Migration in POP1, 10% Migration in POP2, POP3, POP4

5. CONCLUSIONS

In this paper, we have presented a brief study of memetics in NIPD analysing the influence of different factors that are crucial for the game evolution.

We have seen the better performance of the Baldwinian model over the Lamarckian one for the experiments considered. Thus we have selected the Baldwinian model to analyze the influence of several parameters. The memory size proved to be an important factor, since a size of a single play implies too little information and therefore a convergence to non-optimal solutions. Despite increasing this size improves the cooperation, a size too big means that the complexity of the algorithm grows dramatically leading to a worse performance.

We have also seen that a number of players too high impairs the cooperation, and that a higher number of mutations applied to an individual helps the local search and therefore the speed of the process.

We also analyzed the implications of organizing the population spatially in neighborhoods, and how big they

must be. The results obtained showed that cooperation is favored in organized populations, performing better the smallest neighborhoods.

Finally, we have seen how the usage of the Island model can be a good asset when dealing with large populations, but we must pay attention to the values of the migration rate, and the moment when it starts. Migrations should not start too early to avoid transitional states in the search, while a migration rate too high can imply a reduction in the cooperation levels. Future work will consider the use of more complex scenarios for the population network, and the use of different island topologies.

REFERENCES

- [1] Dawkins, R. 1976. *The Selfish Gene*, Oxford University Press.
- [2] Knowles J. and Corne D. 2000. *Memetic Algorithms for Multiobjective Optimization: Issues, Methods and Prospects*.
- [3] Krasnogor N. and Smith J. 2003. "A Tutorial for Competent Memetic Algorithms: Mode, Taxonomy and Design Issues". (Dec.)
- [4] Castillo P. A.; Arenas M. G.; Castellano J. G.; Merelo J. J.; Prieto A.; Rivas V. and Romero G. 2006. "Lamarckian Evolution and the Baldwin Effect in Evolutionary Neural Networks". (March).
- [5] Axelrod R. 1984. *The Evolution of Cooperation*.
- [6] Yao X. and Darwen P. J. 1994. *An Experimental Study of N-Person Iterated Prisoner's Dilemma Games*.
- [7] Nghia Le M.; Soon Ong Y.; Jin Y. and Sendhoff B. 2009. "Lamarckian Memetic Algorithms: Local Optimum and Connectivity Structure Analysis" (Nov.)
- [8] Schweitzer F.; Behera L. and Mühlenbein H. 2002. "Evolution of Cooperation in a Spatial Prisoner's Dilemma". (Nov.)
- [9] Nakamaru, N.; Matsuda, H. and Iwasa, Y. 1997. "The Evolution of Cooperation in a Lattice Structured Population".
- [10] Nowak M. A. and May R. M. 1993. *The Spatial Dilemmas of Evolution*. World Scientific Publishing Company.
- [11] Jovanovic R.; Tuba M. and Simian D. 2010. "Comparison of Different Topologies for Island-Based Multi-Colony Ant Algorithms for the Minimum Weight Vertex Cover Problem". (Jan.)
- [12] Rucinski M.; Izzo D. and Biscani F. 2010. "On the Impact of the Migration Topology on the Island Model". (Apr.)
- [13] Morrison Ronald W.; De Jong Kenneth A. 2002. "Measurement of Population Diversity". Springer-Verlag Berlin Heidelberg
- [14] Garg P. 2009. "A Comparison between Memetic algorithm and Genetic algorithm for the cryptanalysis of Simplified Data Encryption Standard algorithm". (Apr.)

A COMPARATIVE STUDY TO EVOLUTIONARY ALGORITHMS

Eva Volna
Martin Kotyrba
Department of Informatics and Computers
University of Ostrava
70103, Ostrava, Czech Republic
E-mail: eva.volna@osu.cz
E-mail: martin.kotyrba@osu.cz

KEYWORDS

Genetic Algorithms (GA), Simulated Annealing (SA), Differential Evolution (DE), Self Organising Migrating Algorithms (SOMA), Travelling Salesman Problem (TSP).

ABSTRACT

Evolutionary algorithms are general iterative algorithms for combinatorial optimization. The term evolutionary algorithm is used to refer to any probabilistic algorithm whose design is inspired by evolutionary mechanisms found in biological species. These algorithms have been found to be very effective and robust in solving numerous problems from a wide range of application domains. In this paper we perform a comparative study among Genetic Algorithms (GA), Simulated Annealing (SA), Differential Evolution (DE), and Self Organising Migrating Algorithms (SOMA). These algorithms have many similarities, but they also possess distinctive features, mainly in their strategies for searching the solution state space. The four heuristics are applied on the same optimization problem - Travelling Salesman Problem (TSP) and compared with respect to (1) quality of the best solution identified by each heuristic, (2) progress of the search from an initial solution until stopping criteria are met.

INTRODUCTION TO EVOLUTIONARY ALGORITHMS

Evolutionary algorithms (EAs) have many interesting properties and have been widely used in various optimization problems from combinatorial problems such as job shop scheduling to real valued parameter optimization (Back et al. 1997). In computer science, evolutionary computation is a subfield of artificial intelligence (more particularly computational intelligence) that involves combinatorial optimization problems. Evolutionary computation uses iterative progress, such as growth or development in a population. This population is then selected in a guided random search using parallel processing to achieve the desired end. Such processes are often inspired by biological mechanisms of evolution. As evolution can produce highly optimised processes and networks, it has many applications in computer science. Problem solution using evolutionary algorithms is shown in Figure 1.

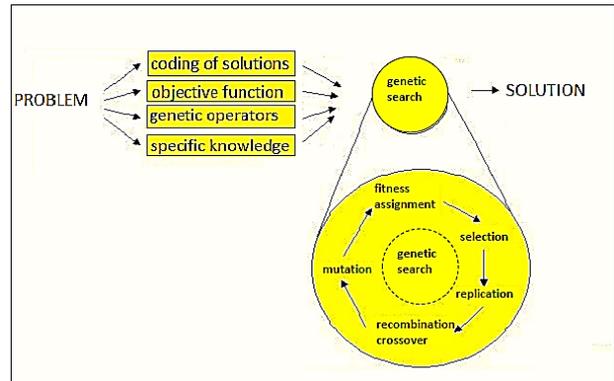


Figure 1: Problem solution using evolutionary algorithms (adapted from <http://jpmc.sourceforge.net>)

Genetic Algorithms

A genetic algorithm is a type of a searching algorithm. It searches a solution space for an optimal solution to a problem. The key characteristic of the genetic algorithm is how the searching is done. The algorithm creates a “population” of possible solutions to the problem and lets them “evolve” over multiple generations to find better and better solutions. The generic form of the genetic algorithm is shown in Figure 2. The items in bold in the algorithm are defined here (Volna 2013).

1. Create a **population** of random candidate solutions named *pop*.
2. Until the algorithm termination conditions are met, do the following:
 - (a) Create an empty population named *new-pop*.
 - (b) While *new-pop* is not full, do the following:
 - i. **Select** two **individuals** at random from *pop* so that individuals, which are more **fit** are more likely to be selected.
 - ii. **Cross-over** the two individuals to produce two new individuals.
 - (c) Let each individual in *new-pop* have a random chance to **mutate**.
 - (d) Replace *pop* with *new-pop*.
3. Select the individual from *pop* with the highest **fitness** as the solution to the problem.

Figure 2: The genetic algorithm

The **population** consists of the collection of candidate solutions that we are considering during the course of the

algorithm. Over the generations of the algorithm, new members are “born” into the population, while others “die” out of the population. A single solution in the population is referred to as an **individual**. The **fitness** of an individual is a measure of how “good” is the solution represented by the individual. The better solution has a higher fitness value – obviously, this is dependent on the problem to be solved. The **selection** process is analogous to the survival of the fittest in the natural world. Individuals are selected for “breeding” (or **cross-over**) based upon their fitness values. The crossover occurs by mingling two solutions together to produce two new individuals. During each generation, there is a small chance for each individual to **mutate**.

Simulated Annealing

Simulated Annealing (SA) was introduced by (Kirkpatrick et al. 1983) for the first time. SA starts off from a randomly selected point. Then a certain number of points is generated in the neighbourhood. The principle of acceptance solution during run of SA is following. If the new cost value is better than the old one new one is accepted immediately. It means that the difference between these two cost values is negative. If the difference is positive (the new cost value is worse than the old one) a number from interval $<0, 1>$ is generated. If it is lower than the probability according to equation (1) the new point is accepted, otherwise the old one continues in the process. This is called Metropolis criterion (Kirkpatrick et al. 1983).

$$p(T) = e^{-\frac{\Delta E}{T}} \quad (1)$$

where $p(T)$ probability of transition for temperature T , ΔE is a difference between cost values of previous and current solution, and T is a current temperature that is a control parameter for cooling schedule.

The algorithm starts with high temperature T , which is decreased in steps. Equation (2) shows standard cooling function.

$$T_{n+1} = \alpha T_n \quad (2)$$

where T_{n+1} is a temperature in the next step, T_n is a temperature in the current step, and α is a coefficient from interval $<0, 1>$.

Differential Evolution

Differential Evolution (DE) is a population-based optimization method that works on real-number-coded individuals (Price 1999). For each individual $\vec{x}_{i,G}$ in the current generation G , DE generates a new trial individual $\vec{x}'_{i,G}$ by adding the weighted difference between two randomly selected individuals $\vec{x}_{r1,G}$ and $\vec{x}_{r2,G}$ to a randomly selected third individual $\vec{x}_{r3,G}$. The resulting individual $\vec{x}'_{i,G}$ is crossed-over with the original individual $\vec{x}_{i,G}$. The fitness of the resulting

individual, referred to as a perturbed vector $\vec{u}_{i,G+1}$, is then compared with the fitness of $\vec{x}_{i,G}$. If the fitness of $\vec{u}_{i,G+1}$ is greater than the fitness of $\vec{x}_{i,G}$, then $\vec{x}_{i,G}$ is replaced with $\vec{u}_{i,G+1}$; otherwise, $\vec{x}_{i,G}$ remains in the population as $\vec{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Figure 3 shows a two-dimensional example that illustrates the different vectors that are used in DE.

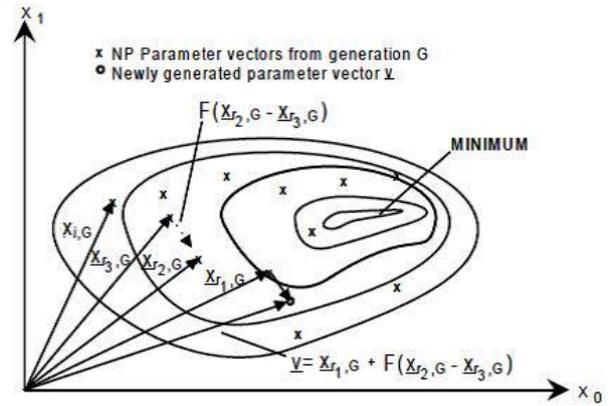


Figure 4: Differential Evolution (Price 1999)

Self-Organising Migrating Algorithm

Self-Organising Migrating Algorithm (SOMA) was developed by prof. Zelinka (Zelinka 2004). SOMA works with groups of individuals (population) whose behaviour can be described as competitive – cooperative strategy. The construction of new population of individuals is not based on evolution principle (two parents produce offspring) but on the behaviour of social group, e.g. a herd of animals looking for food. During one generation, in the case of SOMA this is called migration loop (ML), only the position of individuals in the search space is changed.

In every migration loop the best individual is chosen, i.e. individual with the minimum cost value, which is called leader. An active individual from the population moves in the direction to leader in the search space. At the end of the movement the position of the individual with minimum cost value is chosen. If the cost value of the new position is better than the cost value of an individual from the old population, the new one appears in new population. Otherwise the old one rests there. The movement is described by equation (3).

$$x_{i,j}^{ML+1} = x_{i,j,START}^{ML} + (x_{L,j}^{ML} - x_{i,j,START}^{ML}) \cdot t \cdot PRTVector_j \quad (3)$$

where $x_{i,j}^{ML+1}$ is a value of i -individual's j -parameter, in step t (in the next migration loop $ML + 1$). $x_{i,j,START}^{ML}$ is a value of i -individual's j -parameter that is the $START$ position in the actual migration loop (ML). $x_{L,j}^{ML}$ is a

value of leader's j -parameter in migration loop ML . Step t is from $\langle 0, \text{by Step to, PathLength} \rangle$. $PRTVector$ is a vector of ones and zeros depended on PRT . If random number from interval $\langle 0, 1 \rangle$ is less than PRT , then 1 is saved to $PRTVector$, otherwise 0 is saved to $PRTVector$. There exist four versions of SOMA, but we use version All-To-One in this work because of least time-consuming computing.

THE TRAVELLING SALESMAN PROBLEM

Evolutionary algorithms can be used to solve the travelling salesman problem (TSP), see Figure 4. For those who are unfamiliar with this problem, it can be stated in two ways. Informally, there is a travelling salesman who services some number of cities, including his home city. He needs to travel on a trip such that he starts in his home city, visits every other city exactly once, and returns home. He wants to set up the trip so that it costs him the least amount of money possible. The more formal way of stating the problem casts it as a graph problem. Given a weighted graph with N vertices, find the lowest cost path from some city v that visits every other node exactly once and returns to v . For a more thorough discussion of TSP, see (Garey and Johnson 1979).

The problem with TSP is that it is an NP-complete problem. The only known way to find the answer is to list every possible route and find the one with the lowest cost. Since there are a total of $(N - 1)!$ routes, this quickly becomes intractable for large N . There are approximation algorithms that run in a reasonable time and produce reasonable results – evolutionary algorithms belong to them.

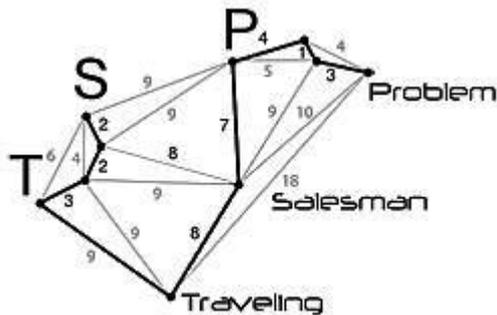


Figure 4: Travelling Salesman Problem (adapted from <http://www.amdusers.com/wiki/tiki-index.php?page=TSP/>)

COMPARATIVE EXPERIMENTAL STUDY

There are selected evolutionary algorithms tested on TSP. To assess the quality of a solution, individual cities are arranged in an orthogonal grid of $N \times N$ points (each point represents one city). Distance between cities is designed as Manhattan distance, i.e. the distance is calculated as an absolute value of a sum of differences between the two coordinates. Manhattan distance d between two cities $P_1=[x_1,y_1]$ a $P_2=[x_2,y_2]$ in plane is calculated as follows (4):

$$d(P_1,P_2) = |x_1 - x_2| + |y_1 - y_2| \quad (4)$$

The advantage of arrangement of cities in the grid lies in the easy comparison, how good is the result. In this arrangement, we are able to easily determine a length of the shortest path, as follows:

If N is even, the shortest path d_{e-min} is calculated as follows (5):

$$d_{e-min} = N^2 \quad (5)$$

If N is odd, the shortest path d_{o-min} is calculated as follows (6):

$$d_{o-min} = N^2 + 1 \quad (6)$$

Fig. 5 represents format of outputs. Individual routes between cities are drawn in different colours. The principle of these colours is the following: (1) black colour represents the shortest distances; (2) pink colour represents distances equal half of the size N of the grid; and (3) blue colour represents all other routes. In the ideal case, the window should contain only black lines. On the top bar of the window (Fig. 5), there are two buttons, which allow switching between results of calculations. Each algorithm was tested on a grid of size 5×5 , 7×7 and 10×10 , which gives 25, 49 and 100 cities. The number of different combinations that can be created in the grid of 10×10 is approximately $9.3 \cdot 10^{157}$. It shows how is hard to find the right solution for permutation problems.

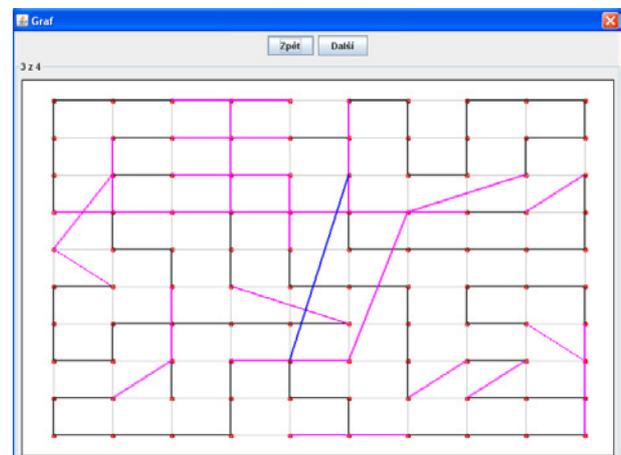


Figure 5: Found route in 10 x 10 grid – 100 cities

Comparative experimental study involves the following evolutionary algorithms:

- Genetic Algorithms (GA).
- Simulated Annealing (SA).
- Differential Evolution (DE).
- Self-Organising Migrating Algorithms (SOMA).

It is necessary to set values of all parameters for each of these algorithms. A minimal number of calculations for each parameters setting was 5000.

Genetic Algorithms

There is a process of reproduction very important in GA. It consists of two parts - the *cross-over* and *mutation*. It

was convenient to separate, for which part of the reproduction was responsible only cross-over and only mutation during experimental study. This was achieved, if we have tested the probability of cross-over, probability of mutation was set to zero and vice versa.

Testing the following parameters *population* and *generation* were merged into one unit, because both parameters determine, how much evaluations were done during calculation. It was an effort to evaluate similar numbers of calculations. For example, if we have reduced a population size to one half, the number of generations was raised twice.

To test CA should be set the following parameters:

Size of population is an integer that specifies a number of chromosomes that occur in a population. *Used values:* 10 – 2000.

Probability of cross-over is a number from the interval $<0; 1>$, which indicates the probability that new individuals will be from two randomly selected parents *Used values:* 0.5 – 1.

Probability of mutation is a number from the interval $<0; 1>$, which indicates the probability with which occurs mutation at individual positions in a chromosome. *Used values:* 0 – 0.1.

Threshold of termination is a number from the interval $<0; 1>$ that indicates how many percent of occurrence of the same chromosome in the population means that a calculation will be terminated. *Used values:* 0.75 – 1.

Maximal number of generations is an integer greater than one indicating how many generations could be done, if some terminal criteria are not fulfilled. *Used values:* 50 – 10000.

Summary:

We could make the following conclusions regarding our experimental study of Genetic Algorithms:

- *Size of population:* > 80 . If the population was smaller, GA started too soon to converge. For example: If the population contains ten individuals and the number of generations has been set to ten thousand, only 463 different combinations were evaluated on average. This means that there is approximately 50 generations before the algorithm was ended.
- *Probability of cross-over:* certainly more than 0.6, but rather in the range of 0.9 - 1 for 25 cities and in the range of 0.8 - 0.9 for 49 cities. If the probability of cross-over was smaller, algorithm quickly converges to a single solution. Although it can be tempered by mutation, but GA always showed worse results at low values of this parameter.
- *Probability of mutation:* strongly depends on the complexity of the problem (i.e. the number of cities). What is the number of cities bigger, the smaller value probability of mutation must be. Its experimental values are the following: 0.01 for 25 cities and 0.004 for 49 cities. If a mutation is too high, the algorithm starts to behave too random, and it is not able to find the minimum. On the

other hand, if a mutation is too small and algorithm quickly converges to a single solution.

- *Threshold of termination* is a very sensitive parameter and depends on all other parameters. If its value was too small, the found route was not usually optimal. If its value was too big, algorithm ran through all generations because it was unable the condition fulfil.
- *Number of generations* should be more than 200 for 25 cities and more than 2000 for 49 cities. If a number of generations were smaller, GA had not a sufficient number of cycles for an evolution of its population, because they were still found unsuitable individuals in the population, who were involved in reproduction.
- Each calculation of GA should be run several times to use the average results. GA was often able to find optimal solutions even if parameters have been set worse. On the contrary, even if parameters have been set ideally, GA was not always able to find the optimal solution.

Simulated Annealing

To test SA should be set the following parameters:

Initial temperature is an integer greater than one, which sets the temperature of a beginning of an annealing process. *Used values:* usually 0.01 - 1000 (e.g. $1000 \times \text{final temperature}$).

Final temperature is a number greater than zero, which represents the final temperature of a termination of a calculation. *Used values:* 0.00001 – 1.

Temperature reduction factor is a number from the interval $<0; 1>$ that indicates about how many percent is the actual temperature cooled at each cycle. *Used values:* 0.01 – 0.999.

Number of iterations at a given temperature is an integer greater than zero that represents how many times is algorithms repeated at a given temperature. *Used values:* 1 – 17200.

Summary:

We could make the following conclusions regarding our experimental study of Simulated Annealing:

- *Temperature reduction factor:* > 0.6 . If the value is smaller, the temperature cools too quickly. Consequently, the big leaps occur in the probability of acceptance of worse solutions. Recommended values are about 0.99 - 0.999. It is appropriate to the probability of acceptance of worse solutions has decreased slowly.
- *Number of iterations at a given temperature:* > 1 . If a large number of iterations is set to be tested a large number of routes before the temperature cools. If a temperature reduction factor is set near one then temperature cools slowly leading to test a larger number of possibilities. Therefore, these parameters were tested together, and they were adjusted so that the number of tested routes was similar.

- *Initial temperature* should be from the interval $\langle 0.1; 50 \rangle$. For higher temperatures, the probability of acceptance of worse solutions is too high. In contrast, the algorithm often gets stuck in a local minimum for smaller temperature.
- Each calculation of SA should be run several times (see GA).

Differential Evolution

To test DE should be set the following parameters:

Threshold of cross-over is a real number from interval $\langle 0, 1 \rangle$, indicating how likely it will be placed at the position an element of noise or target vector. *Used values:* 0.05 – 0.9.

Size of population is the number of individuals in a population (see GA). *Used values:* 10 – 800.

Mutation constant is a real number and indicates how much will be multiplied the vector during a mutation. *Used values:* 0 – 2.

Number of generations see GA. *Used values:* 125 – 10000.

Summary:

We could make the following conclusions regarding our experimental study of Differential Evolution:

- The greater the complexity of the problem is, the smaller have to be set *threshold of cross-over*: range of 0.2 – 0.3 for 25 cities and in the range of 0.1 – 0.1 for 49 cities. The real values are usually set higher than the above recommended values in order to DE not quickly converge to a single solution. It could be used especially when it is set greater number of generations.
- Regarding setting a *mutation constant*, there were not recognized achieved significant differences in results depending on the setting of this parameter. In this case, it is possible to say, that the parameter is not significant in the course of DE.
- To prevent a stagnation process, it is necessary to do the following: either to set up a large number of individuals in a population, or threshold of cross-over should be higher.
- *Number of generations* should be higher then a complexity of the problem is. Simultaneously, it is suitable to set a size of the population size at higher value, because the parameter has proved as decisive regarding stagnation.
- Each calculation of DE should be run several times (see GA).

Self-Organising Migrating Algorithms

To test SOMA should be set the following parameters:

Migration is a parameter, which equals to the parameter maximal number of generations in GA. *Used values:* 20 – 10000.

Minimal diversity means a terminal distance between the best and the worst individual in a given population. It is a terminating parameter (integer). *Used values:* 0 – 18 (30).

Path Length is a real number from the interval $\langle 1; 5 \rangle$. It determines how far an individual goes on its way toward the best individual. *Used values:* 0.33 – 10.

Step is a real number from the interval $\langle 0.11 - Path Length \rangle$. It determines how many times individual stops during its path. *Used values:* 0.033 – 0.9.

Perturbation (PRT) is a parameter that replaces the mutation from GA. *Used values:* 0 – 1.

Size of population is the number of individuals in a population (see GA). *Used values:* 5 – 500.

Summary:

We could make the following conclusions regarding our experimental study of SOMA:

- *Perturbation* is set to small values ranging $\langle 0, 0.1 \rangle$. The larger the dimension of the problem is, the smaller has to be the value of the parameter.
- *Migration* > 1000 . The larger dimension of the problem is, the greater has to be the value of the parameter. *Migration* is able to be replaced with parameters *step* and *path length*.
- It is not appropriate to set the parameter *path length* so that it would be a multiple of the parameter *step*. In such case, there is often a quick convergence caused by the fact that each individual is on its way stops exactly at the place where is situated the best individual.
- *Minimal diversity*. It is evident, if a parameter minimal diversity is set too high the algorithm terminates before an appropriate solution is found. On the other hand, if the parameter is set too low, the calculation will never be stopped by this parameter. During the experiments, it was shown that if the algorithm terminated by this parameter, the achieved solutions has worse values. It is appropriate to set the value of this parameter to zero.
- The biggest influence on the quality of the solution had *population*. A large population clearly reached the best results for 25 cities. Large population was not enough and proved to be the weakest parameter for 49 cities. This is a phenomenon that was observed in all tested algorithms.

OUTCOMES

Finally, all of these algorithms were tested in 10×10 grid, i.e. 100 cities. Due to the complexity of the problem, each algorithm was run only twice. There are parameters settings at single algorithms in Table 1.

Outputs of single algorithms are presented in the following graph (Fig. 6). From this graph it is clear that SA achieved the best results when it found a way which indicates the third best possible value. According to formula (5) the known best way is the following: $d_{e_{min}} = 10^2 = 100$. Fig. 7 shows that SA has found the ways, where $d = 104$.

Table 1: Parameters settings at single algorithms.

GA	SA
Size of population: 400 Probability of cross-over: 0.9 Probability of mutation: 0.001 Threshold of termination: 0.5 Max. number of generations: 60000	Initial temperature: 50 Final temperature: 0.0001 Temperature reduction factor: 0.999 Number of iterations at a given temperature: 1000
DE	SOMA
Size of population: 3000 Number of generations: 40000 Threshold of cross-over: 0.1 Mutation constant: 0.8	Perturbation: 0.01 Migration: 30000. Minimal diversity: 2 Path Length: 3 Step: 0.11 Size of population: 400

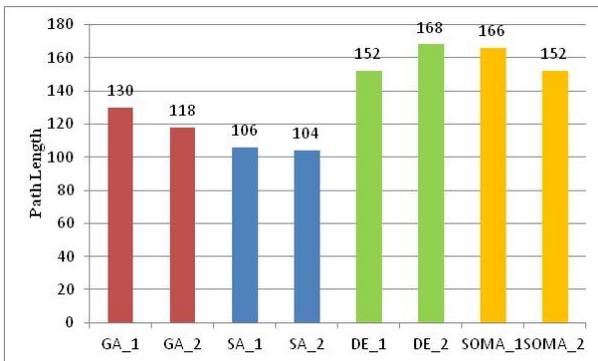


Figure 6: Comparative results of single algorithms with 100 cities

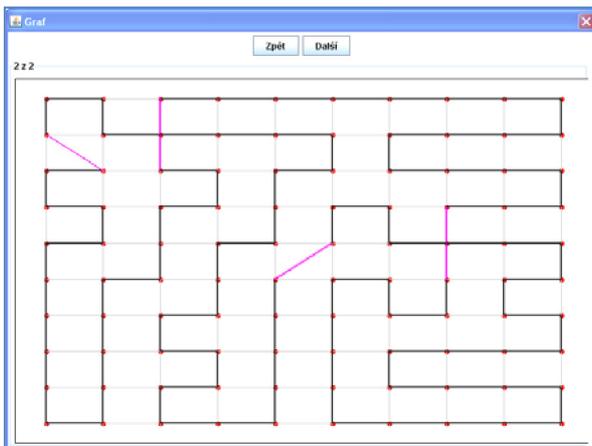


Figure 7: Simulated Annealing – 100 cities

CONCLUSION

In this paper, an experimental comparative study of four popular approximation algorithms (GA, SA, DE, and SOMA) is presented for travelling salesman problem. All four heuristics assume and exploit regularities present within the search space, i.e., search spaces where good solutions have higher probabilities of leading to better solutions. All four tested algorithms have been found to be effective and robust on problems where some measure of progress can be shown. These algorithms discussed incorporate domain specific knowledge to dictate the search strategy. The principle

deference among them is how and where domain-specific knowledge is used. In this work our intention has been to study the behaviour of the four heuristics in solving a hard combinatorial problem, and not to demonstrate the superiority of one algorithm over the other over all problem domains. Each one of them has its merits. Actually it would be unwise to generalize the results reported here over all classes of problems. To solve such question would require at least that similar experiments on other category of problems be performed. Such experiments are the subject of our future work.

ACKNOWLEDGMENT

The research described here has been financially supported by University of Ostrava grant SGS/PřF/2014.

REFERENCES

- Back, T., Hammel U. and Schwefel H.-P. 1997. Evolutionary Computation: Comments on the History and Current State. *IEEE Trans. on Evolutionary Computation*, pp. 3-17.
- Garey, M. and Johnson, D.S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company.
- Hrabal, R. 2010. NP-problems solving via evolutionary algorithms. Master thesis (in Czech). University of Ostrava. 66 p.
- Kirkpatrick S., Gelatt C. D., Vecchi M. P. 1983. Optimization by Simulated Annealing, *Science*, 13, Volume 220, Number 4598, pp. 671 – 680
- Price, K. 1999. An Introduction to Differential Evolution. In: D. Corne, M. Dorigo and F. Glover (eds.) *New Ideas in Optimization*. London: McGraw-Hill, pp. 79–108.
- Volná, E. 2013. *Introduction to Soft Computing*. bookboon.com Ltd. ISBN: 978-87-403-0391-9 (available from <http://bookboon.com/en/introduction-to-soft-computing-ebook>)
- Zelinka I. 2004. SOMA – Self Organizing Migrating Algorithm, In: Babu, B.V. Onwubolu G. (eds), *New Optimization Techniques in Engineering*. Springer-Verlag, ISBN 3-540-20167X



EVA VOLNÁ is an associate professor at the Department of Computer Science at University of Ostrava, Czech Republic. Her interests include artificial intelligence, artificial neural networks, evolutionary algorithms, and cognitive science. She is an author of more than 50 papers in technical journals and proceedings of conferences.



MARTIN KOTYRBA is an assistant professor at the Department of Computer Science at University of Ostrava, Czech Republic. His interests include artificial intelligence, formal logic, soft computing methods and fractals. He is an author of more than 30 papers in proceedings of conferences.

COMPARISON OF MODERN CLUSTERING ALGORITHMS FOR TWO-DIMENSIONAL DATA

¹Martin Kotyrba, ¹Eva Volna, ²Zuzana Kominkova Oplatkova

¹Department of Informatics and Computers
University of Ostrava, 70103, Ostrava, Czech Republic
martin.kotyrba@osu.cz, eva.volna@osu.cz

²Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
kominkovaoplatkova@fai.utb.cz

KEYWORDS

Cluster analysis, K-Means, Self Organising Map Algorithm, DBSCAN.

ABSTRACT

Cluster analysis or clustering is a task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is the main task of exploratory data mining and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The topic of this paper is modern methods of clustering. The paper describes the theory needed to understand the principle of clustering and descriptions of algorithms used with clustering, followed by a comparison of the chosen methods.

INTRODUCTION TO CLUSTER ANALYSIS

Cluster analysis itself is not one specific algorithm, but a general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to find them efficiently. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, density threshold, or number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection

of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy and typological analysis. Subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals. In this paper we will compare three different algorithms in an experimental study.

MODERN CLUSTERING METHODS

There are some well used clustering algorithms out there; one of them is the famous CLARANS. Other methods are K - means, K-medoid, Hierarchical Clustering and Self-Organized Maps. Nevertheless, none of these algorithms can handle all these three mentioned problems in a good way. This report will not discuss these methods but focus on the DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm, which introduces solutions to these problems.

DBSCAN

It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. OPTICS can be seen as generalization of DBSCAN to multiple ranges, effectively replacing the ϵ parameter with a maximum search radius. DBSCAN's definition of a cluster is based on the notion of density reachability. Basically, a point q is directly density-reachable from a point p if it is not farther away than a given distance ϵ (i.e., it is part of its ϵ -neighborhood) and if p is surrounded by sufficiently many points such that one may consider p and q to be part of a cluster. Q is called density-reachable (note the distinction from "directly density-reachable") from p if there is a sequence $p_1 \dots p_n$ of points with $p_1 = p$ and $p_n = q$ where each p_{i+1} is directly density-reachable from p_i .

Note that the relation of density-reachable is not symmetric. Q might lie on the edge of a cluster, having insufficiently many neighbors to count as dense itself. This would halt the process of finding a path that stops with the first non-dense point. By contrast, starting the process with p would lead to q (though the process would halt there, q being the first non-dense point). Due to this asymmetry, the notion of density-connected is introduced: two points p and q are density-connected if there is a point 0 such that both p and q are density-reachable from 0 . Density-connectedness is symmetric.

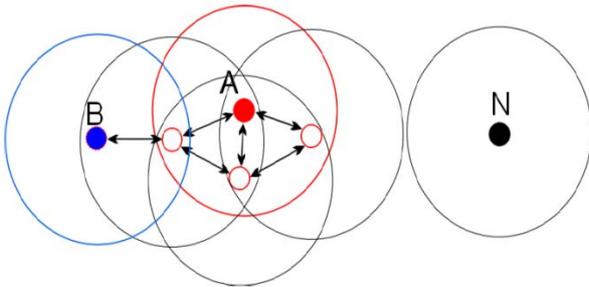


Figure 1: DBSCAN

A cluster, which is a subset of the points of the database, satisfies two properties:

1. All points within the cluster are mutually density-connected.
2. If a point is density-connected to any point of the cluster, it is part of the cluster as well.

DBSCAN (Fig.1) requires two parameters: ϵ (eps) and the minimum number of points required to form a cluster ($minPts$). It starts with an arbitrary starting point that has not been visited. This point's ϵ -neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ϵ -environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ϵ -neighborhood is also part of that cluster.

The algorithm DBSCAN:

1. Select the object from the set of
2. Match all reachable points from the selected item, if applicable, the ϵ -neighborhood contains at least $MinPts$, will form a new cluster.
3. Find objects directly reachable from these cores may be joining clusters.
4. Cease at the moment when none of the remaining objects can no longer be added to any cluster.

Figure 2: The DBSCAN algorithm

Hence, all points that are found within the ϵ -neighborhood are added, as is their own ϵ -neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is

retrieved and processed, leading to the discovery of a further cluster or noise. DBSCAN disadvantage of the method is its sensitivity to parameter settings and ϵ $MinPts$, the main advantage lies in the ability to distinguish clusters of different shapes, resistance to remote objects, and especially the detection of clusters. In Fig. 2 you can see basic steps of DBSCAN algorithm.

K-Means Clustering

K-means (Fig.3) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

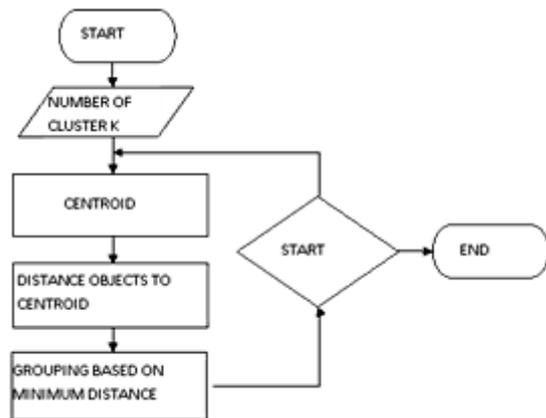


Figure 3: K-Means

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function (1)

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

Where $\|x_i^{(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster c_j is an indicator of the distance of the n data points from their respective cluster centres. The algorithm (Fig. 4) is composed of the following steps:

The algorithm K-Means

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Figure 4: The K-Means algorithm

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many problem domains. As we are going to see, it is a good candidate for extension to work with fuzzy feature vectors.

SOM

The Kohonen Self-Organizing Feature Map (SOFM or SOM) is a clustering and data visualization technique based on a neural network viewpoint. As with other types of centroid-based clustering, the goal of SOM is to find a set of centroids (reference or codebook vector in SOM terminology) and to assign each object in the data set to the centroid that provides the best approximation of that object. In neural network terminology, there is one neuron associated with each centroid. As with incremental K-means, data objects are processed one at a time and the closest centroid is updated. Unlike K-means, SOM impose a topographic ordering on the centroids and nearby centroids are also updated. The processing of points continues until some predetermined limit is reached or the centroids are not changing very much. The final output of the SOM technique is a set of centroids that implicitly define clusters. Each cluster consist of the points closest to a particular centroid. SOM is a clustering technique that enforces neighborhood relationships on the resulting cluster centroids. Because of this, clusters that are neighbors are more related to one another than clusters that are not. Such relationships facilitate the interpretation and visualization of the clustering results. Indeed, this aspect of SOM has been exploited in many areas, such as visualizing Web documents or gene array data.

A distinguishing feature of SOM is that it imposes a topographic (spatial) organization on the centroids (neurons). An example of a two-dimensional SOM in which the centroids are represented by nodes that are organized in a rectangular lattice. Each centroid is assigned a pair of coordinates(i,j). Sometimes, such a network is drawn with links between adjacent nodes, but can be misleading because the influence of one centroid

on another is via a neighborhood that is defined in terms of coordinates, not links. There are many types of SOM neural networks, but it will be focus on to two-dimensional SOMs with a rectangular or hexagonal organization of the centroids.

Even though SOM is similar to K-means, there is a fundamental difference. Centroids used in SOM have a predetermined topographic ordering relationship. During the training process, SOM uses each data point to update the closest centroid and centroids that are nearby in the topographic ordering. In this way, SOM produces an ordered set of centroids for any given data set. In other words, the centroids that are close to each other in the SOM grid are more closely related to each other than to the centroids that are farther away. Because of this constraint, the centroids of a two-dimensional SOM can be viewed as lying on a two-dimensional surface that tries to fit the n-dimensional data as well as possible. The SOM centroids can also be thought of as the result of a nonlinear regression with respect to the data points. At a high level, clustering using the SOM technique consists of the steps described in Algorithm which you can see in Fig.5.

The algorithm SOM:

1. Determine the number of clusters.
2. Initialize the cluster centers.
3. Compute partitioning for data.
4. Compute (update) cluster centers.
5. If the partitioning is unchanged (or the algorithm has converged), stop; otherwise, return to step 3.

Figure 5: The SOM algorithm

EVALUATION CRITERIA

On the basis of the experiments these parameters were chosen in individual methods: the ability to determine the number of clusters, sensitivity to outlying values, ability to distinguish clusters of arbitrary shape, sensitivity settings from the user. Table 1 shows a comparison of different methods based on the monitored parameters

Table 1: Parameter setting for individual algorithms

Name of method	Arbitrary shape of clusters	Sensitivity settings	Sensitivity to outlying objects	Determination of the number of clusters
DBSCAN	yes	yes	no	yes
SOM	yes	yes	no	yes
K-Means	no	no	yes	no

Based on this comparison, the following was chosen for implementation of these methods: method of k-means, SOM method and method DBSCAN. The SOM method was chosen as a representative model-based methods, DBSCAN is selected as the representative method based on density. The K-means method was chosen to be

implemented due to comparison with these modern methods. Another reason for its selection was illustrated by the inability of the clustering method to distinguish clusters of arbitrary shape. An equally important reason is that it is one of the most used clustering methods.

EXPERIMENTAL STUDY

In an experimental study, the methods were compared on the three data sets of two-dimensional data. The experimental data set contains three clusters arranged in shape. For demonstration we present experiments on only one data set. This data set has been selected for implementation because of the presentation capabilities of clustering methods to deal with the cluster complex shape such as used in a spiral. In the results for each method, there is a table containing the column number of the cluster, which indicates the number of clusters where appropriate. The column number of objects in a cluster that indicates how many objects are assigned to a cluster, the column contains the percentage distribution of the percentage of the size of the cluster to the total number of objects, see Tab.2.

Table 2: Data sets represent clusters arranged in a two-dimensional spiral

Number of objects	529
Clusters	3
Dimension	2
Noise	0%
The he shape of clusters	spiral

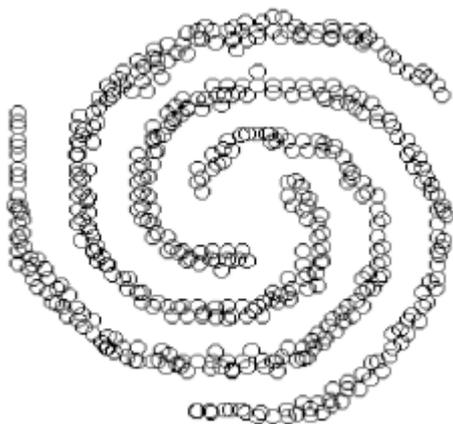


Figure 6: The shape of used spiral

The result of the using a method K-Means when the number of clusters $k = 3$, see Tab.3 and Fig.7.

Table 3: Results for K-Means method

Cluster	Number of objects in cluster	Percentage distribution
0	170	32%
1	171	32%
2	188	36%

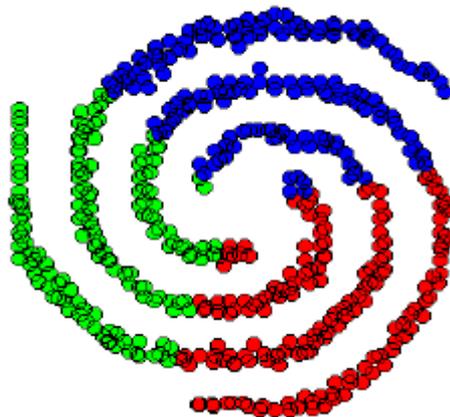


Figure 7: Results for K-Means methods

The result of the using the DBSCAN method when was following set: MinPts=6 and $\epsilon=20$, see Tab.4 and Fig. 8.

Table 4: Results for DBSCAN method

Cluster	Number of objects in cluster	Percentage distribution
0	196	37%
1	168	32%
2	165	31%

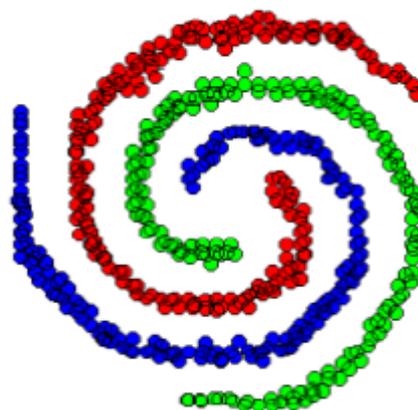


Figure 8: Results for DBSCAN methods

The result of using the SOM method when the output layer was 8×8 neurons and $\alpha=1.0$. SOM method using the size of the output layer 8×8 where was formed 62 clumps in Figure X is visible placement of neurons in the area of clusters with respect to their shape. A neuron is indicated by a black circle with the numerical designation of its position in the output layer, see Tab.5 and Fig. 9.

Table 5: The first column represents neuron, the second number of objects in cluster and the third percentage distribution.

n _{0,0}	8	2%	n _{2,0}	7	1%	n _{4,0}	8	2%	n _{6,0}	7	1%
n _{0,1}	5	1%	n _{2,1}	10	2%	n _{4,1}	7	1%	n _{6,1}	4	1%
n _{0,2}	6	1%	n _{2,2}	12	2%	n _{4,2}	19	4%	n _{6,2}	8	2%
n _{0,3}	8	2%	n _{2,3}	14	3%	n _{4,3}	5	1%	n _{6,3}	7	1%
n _{0,4}	10	2%	n _{2,4}	10	2%	n _{4,4}	8	2%	n _{6,4}	6	1%
n _{0,5}	10	2%	n _{2,5}	7	1%	n _{4,5}	7	1%	n _{6,5}	10	2%
n _{0,6}	8	2%	n _{2,6}	8	2%	n _{4,6}	0	0%	n _{6,6}	8	2%
n _{0,7}	7	1%	n _{2,7}	10	2%	n _{4,7}	12	2%	n _{6,7}	10	2%
n _{1,0}	7	1%	n _{3,0}	9	2%	n _{5,0}	8	2%	n _{7,0}	4	1%
n _{1,1}	10	2%	n _{3,1}	10	2%	n _{5,1}	9	2%	n _{7,1}	7	1%
n _{1,2}	9	2%	n _{3,2}	14	3%	n _{5,2}	7	1%	n _{7,2}	10	2%
n _{1,3}	10	2%	n _{3,3}	0	0%	n _{5,3}	9	2%	n _{7,3}	10	2%
n _{1,4}	8	2%	n _{3,4}	9	2%	n _{5,4}	7	1%	n _{7,4}	8	2%
n _{1,5}	8	2%	n _{3,5}	7	1%	n _{5,5}	3	1%	n _{7,5}	8	2%
n _{1,6}	9	2%	n _{3,6}	12	2%	n _{5,6}	8	2%	n _{7,6}	9	2%
n _{1,7}	10	2%	n _{3,7}	7	1%	n _{5,7}	8	2%	n _{7,7}	9	2%

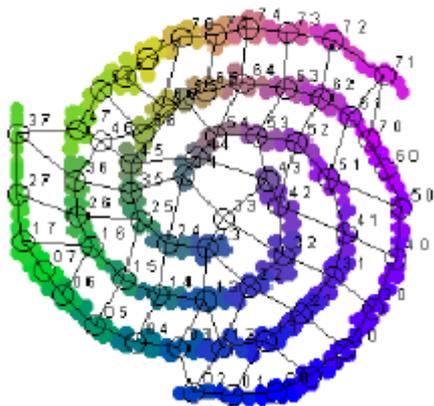


Figure 9: Results for SOM method

The data set in spiral shape is a clearly evident inability of K-Means method to take into account the shape of cluster. The DBSCAN method has solved this problem very well, discovered all the clusters taking into account their shape. The SOM method in the grid smaller than 8x8 could not take into account the shape of a spiral occurred linking individual shoulder, but Figure 9 shows a 8x8 grid settings distinguish 62 clusters, the largest had 19 objects, the method create a higher number of smaller clusters (Sefar, 2013).

CONCLUSION

Our experiments have shown that the best method for two-dimensional data clustering is DBSCAN. The method was able to distinguish all clusters correctly, its drawback lies in the possibility of a more difficult setting of initial parameters, which do not provide the desired result immediately. This quality of the DBSCAN method can be a problem when we do not have information on the number of clusters in the data

set. The SOM method suffers from a similar drawback as DBSCAN. It is necessary to adequately set the number of neurons in the output layer. A low number can result in grid over twisting or in representing more objects by one neuron, which results in failing to distinguish the shape of the cluster. The k-means method, as presumed, wasn't able to distinguish clusters of arbitrary shape and deal with remote objects. Using the k-means method seems to be favorable if we do not need to take into consideration the cluster shape and we have information on their number, or we require their exact number.

ACKNOWLEDGEMENT

The research described here has been financially supported by University of Ostrava grant SGS/PřF/2014 and by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

REFERENCES

- MacQueen, J. B. 1967, Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297
- Sefar, S. 2013, Modern clustering methods, Bachelor thesis, University of Ostrava.
- Moore, A. 2005, K-means and Hierarchical Clustering - Tutorial Slides <http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html>
- Cabanes, G. & Bennani, Y. 2007. A simultaneous two-level clustering algorithm for automatic model selection. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA'07)* Cincinnati, Ohio, USA.
- Kohonen, T. 2001. *Self-Organizing Maps*. Berlin: Springer-Verlag.
- Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., Müller-Gorman, I., Zimek, A. 2007. Detection and Visualization of Subspace Cluster Hierarchies. *LNCS: Advances in Databases: Concepts, Systems and Applications*. Lecture Notes in Computer Science 4443: 152–163.
- Chakraborty, S., Nagwani, N.K., Dey, L., 2011, Performance Comparison of Incremental K-means and DBSCAN algorithms. *International Journal of Computers Applications* (097-8887), Volume 27 – No.11, August.



EVA VOLNA is an associate professor at the University of Ostrava. Her interests include artificial intelligence, artificial neural networks, evolutionary algorithms, and cognitive science. She is an author of more than 50 papers in technical journals and conference proceedings.



MARTIN KOTYRBA His interests include artificial intelligence, formal logic, soft computing methods and fractals. He is an author of more than 30 papers in conference proceedings.



ZUZANA KOMINKOVA

OPLATKOVA is an associate professor at Tomas Bata University in Zlin. Her research interests include artificial intelligence, soft computing, evolutionary techniques, symbolic regression, neural networks. She is an author of around 100 papers in journals, book chapters and conference proceedings. Her e-mail address is: oplatkova@fai.utb.cz

REUSABLE REINFORCEMENT LEARNING FOR MODULAR SELF MOTIVATED AGENTS

Jaroslav Vítků, Pavel Nahodil
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Technická 2, 16627, Prague 6, Czech Rep.
Email: vitkujar@fel.cvut.cz, nahodil@fel.cvut.cz

KEYWORDS

Agent, Architecture, Artificial Life, Creature, Behaviour, Hybrid, Neural Networks, Evolution.

ABSTRACT

Presented topic is from the research fields called Artificial Life and Artificial Intelligence (AI). In this paper, there is presented novel approach to designing agent architectures with its requirements. The approach is inspired by inherited modularity of biological brains and agent architectures are represented here as set of given reusable modules connected into a particular topology. This paper presents design of two particular modules for future use in more complex architectures. The modules are used for implementing model-free motivation-driven Reinforcement Learning (RL). First, the novel framework for these architectures is described together with a used simulator. Then, the design of two new reusable domain-independent components of agent architectures is described. Finally, experimental validation of these new components and their future use is mentioned.

INTRODUCTION

This paper deals with design of agent architectures in the domain of Artificial Life (ALife), which is often inspired in *ethology*. In ethology, the emphasis is put on agents behaviour. Observing agents behaviour and determining its origins (decomposition of problem) is one possible source of inspiration for architecture design (called "top-down" approach). The other possible approach (called "bottom-up") is in connecting small systems (capable of simple behaviour) into larger architectures. This way of designing intelligent systems is often called *connectionism*. Such connectionist models may have promising future with new, more detailed models of neurons Maass (1996); Izhikevich (2003) together with emerging specialized hardware Thomas and Luk (2009) for them. Each approach has own advantages and drawbacks. Our focus is aimed more towards combining the two above together into new, hybrid systems. These hybrid architectures partly employ ethological principles and partly connectionist ones. Our approach focuses on

reusability of particular components of agent architectures. This paper describes two reusable modules, which can be freely used in variety of modular architectures.

The first chapter of this paper describes our novel framework for representation and design of hybrid agent architectures, its goals and requirements. The second chapter describes theory and implementation of two new modules which implement domain-independent and model-free Reinforcement Learning (RL). The chapter Experiments describes experiments simple architecture used for verification of these modules. Finally, results of experiments are evaluated and future use of these modules in automatic design of agent architectures is mentioned.

HYBRID ARTIFICIAL NEURAL NETWORK SYSTEMS FRAMEWORK

Rather than designing one particular architecture suitable for a particular task, our research focuses on modular systems Auda and Kamel (1999) where each module can be reused in various architectures. An example of typical reusable domain independent module can be seen the Categorizing and Learning Module (CALM) Murre et al. (1989).

Neural Module

For this purpose, the framework called Hybrid Artificial Neural Network Systems (HANNNS) is described. Its main goal is in unification of representation of particular modules, so that these modules can be seamlessly connected into bigger systems. Particular sub-systems use for communication the same methods as ANNs and are defined as "Neural Modules". Each Neural Module can have Multiple Inputs/Multiple Outputs (MIMO), either real-valued or spiking type. Neural Module can implement theoretically any component of agent architecture: sensory systems, decision-making modules or actuators. Scheme of Neural Module can be seen in the Fig.1, where particular components are explained in this section in more detail.

Prosperity Measure: The other main goal of this framework is enabling automatic design of agent architectures by means of Evolutionary Algorithms (EAs). Here, similar methods to neuro-evolution Fekiac et al. (2011) can be used. While omitting multi-objective

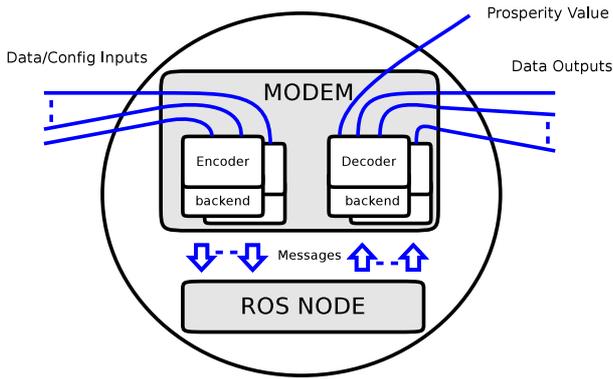


Fig. 1. Scheme of Neural Module with three data and configuration inputs, three outputs, one "prosperity" output and arbitrary inner structure. An encapsulated algorithm can be implemented by means of Robotic Operating System (ROS). The prosperity output represents subjective heuristics telling how well the algorithm performs.

optimization techniques Deb (2011), the evolutionary approach requires single measure of quality of a solution (architecture) represented by an individual. This measure is called the *Fitness*.

Often, the quality of a solution as a whole is evaluated. However architectures represented by HANNS can be composed of highly heterogeneous modules (from simple functions to complex algorithms). In order to give some insight whether a particular module is used efficiently in a given architecture, it is suitable to have some measure representing this. Therefore this framework defines the *Prosperity*. The Prosperity is similar to the Fitness, but represents heuristics which tells "how well the particular Neural Module probably works". The value should be from the interval $\langle 0, 1 \rangle$, where 0 represents the worst, and 1 the best performance. This value is "subjective" and is provided by the Module itself during the simulation.

Data Connections and Configuration Connections: Neural Module is purposed to implement various algorithms. Despite the fact that such algorithms should be domain-independent, some configuration is needed in many cases. In most cases, the values of parameters are constant during the simulation. Therefore the HANNS framework distinguishes data and configuration connections. This can be used by specialized EA algorithms for separate searching for connections and parameters (such as Kordík (2006)) or to use predefined parameters and search only for data connections.

The Simulator NengoROS

In order to create and use as reusable modules as possible, the simulator NengoROS (available at [url-http://nengoros.wordpress.com/](http://nengoros.wordpress.com/)) was created. It combines simulator of large-scale neural networks (based on Neural Engineering Framework (NEF) Eliasmith and Anderson (2003)) called Nengo (available at nengo.ca) and the Robotic Operating System (available at <http://www.ros.org/>). The ROS is decentralized infrastructure

based on nodes Quigley et al. (2009), which communicate by means of messages over the TCP/IP protocol. In the ROS, each node is separated process (several programming languages supported so far), by connecting several nodes together, a network-like structure can be created. Particular algorithms can be used (or implemented) as ROS Nodes, where each ROS node has own Jython interface which connects it into the NengoROS simulator. The Jython interface defines modem, a ROS node which translates ROS messages into the Nengo data, see Fig.1.

THEORETICAL BACKGROUND

When compared to the knowledge-based AI and to connectionism approaches, several types of RL algorithms have several advantages. Compared to supervised-learning ANNs, these algorithms do not require learning by examples. And compared to planning systems, they do not require even a model of the environment. RL is based only on rewards received as a result of some action executed. This makes RL algorithms suitable for unknown environments and also usable in the HANNS framework. For the integration, the type of RL, called Q-Learning was chosen.

Q-Learning Algorithm

The Q-Learning algorithm is suitable for online learning without need of environment model - it is model-free approach. During the learning, the algorithm updates the action-value function Q , which represents mapping set of agent's actions A and set of all admissible environment states S to real values according to the equation (1).

$$Q : A \times S \rightarrow \mathbb{R} \quad (1)$$

Values in the matrix $Q(s, a)$ then define the benefit of each action in a given actual state. When exploiting the knowledge learned the by Q-Learning algorithm, the best action (with the highest value in the matrix) can be selected at each step for obtaining best known policy in a given situation. At each step of the algorithm, values in the $Q(s, a)$ matrix are updated according to the equation (2):

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha [r_t + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \quad (2)$$

Here, the $s_t \in S$ is a previous state of the environment, $a_t \in A$ is action which was just executed, r_t is reward received at a result to the action a_t , the current time step t , $\max_a Q(s_{t+1}, a_{t+1})$ is the action with the highest utility value in the current state. There are the following algorithm parameters: $\gamma \in \langle 0, 1 \rangle$ is a forgetting factor and $\alpha \in \langle 0, 1 \rangle$ is a learning rate, for more information see Vítků (2011).

The scheme of Q-Learning system and the principle of it's function is depicted in the figure 2. The Stochastic

Return Predictor (SRP) is composed of Q-Learning algorithm and Action Selection Method (ASM). The ASM selects the action and executes it, RL algorithm observes the reinforcements received and updates the value of the Q function for the previous state according to the eq. (2).

The algorithm can be further improved by the *Eligibility Traces*. When using the Eligibility traces, the algorithm updates values of several state-action pairs at one step Vítků (2011). The parameter λ defines how much are previous states updated. Correct estimation of the λ can greatly improve the speed of learning, but also can cause oscillations in learning. This modification is also called *Q-Lambda* algorithm.

Action Selection Method: Actions to be executed by the agent are selected by the ϵ -Greedy Action Selection Method (ASM). The ϵ parameter defines amount of randomization: with the probability of ϵ , a random action is selected and with the probability of $1 - \epsilon$ the Greedy action is taken. This helps the agent escape from the local extreme and encourages exploration of new states.

Motivation Source

As seen in the previous chapter, the amount of randomization (exploration of the state-space) can be chosen by varying the ϵ value. This can be taken further and enable the agent architecture to set the ϵ parameter dynamically during the simulation, based on the current situation. If there is a free time, living animals tend to play/explore the surrounding environment and therefore gain new knowledge. On the other hand, if the situation requires fast exploitation of current knowledge, it is not suitable to explore. In many systems, there is need to choose good tradeoff between *exploration* and *knowledge exploitation*.

In the past, our research team solved this for example by defining agents physiology Kadleček (2008); Kadleček and Nahodil (2001); Kadleček and Nahodil (2008). The physiological variable can represent e.g. need for water. In time, as the value of variable (e.g. amount of water in body) decreases, the need for correcting this state (drinking) increases. After drinking, the variable is returned towards the optimal condition and the motivation decreases.

Exactly this purpose has the *Motivation Source*. The simplest case can include one linearly decaying physiological variable, where the amount of motivation inversely depends on a value of the physiological variable. Here, a more natural definition is used: the physiological variable decays linearly each time step t with predefined *decay*:

$$V_{t+1} = V_t - \text{decay}, \quad (3)$$

but the amount of motivation is determined by applying the sigmoid to the inverse value of physiological variable V . The resulting amount of Motivation M at time t is:

$$M_t = \frac{1}{1 + e^{\min + (\max - \min) * (1 - V_t)}}, \quad (4)$$

where \min and \max parameters are chosen, so that value of the variable $V_t = 0$ roughly corresponds to the motivation of $M_t = 1$.

DESIGN OF NEURAL MODULES

This section describes the design of two reusable Neural Module that can be used for model-free learning in modular agent architectures. This section briefly describes design and implementation of these modules. The first module implements the *Q-Lambda* algorithm together with the ASM. The second Module implements the *Physiological State Space*, which can serve as a source of motivation in agent architecture Kadleček (2008).

Q-Learning Module Design

Several design requirements have to be met in order to successfully implement the Q-Lambda algorithm in the Neural Module. First, the typical use-case and main requirements for such a Neural Module in the HANNS framework will be described. The requirements for integration of the Q-Lambda algorithm into the Neural Module are described in the following sections.

Representing the Inputs and Outputs: Neural Modules in the HANNS communicates by vector of real values on the interval $\langle 0, 1 \rangle$. Since the module should be as compatible with classical ANN paradigms as possible, the encoding of input/output values is selected *1ofN*. In case of actions, only the currently selected one has non-zero value on its output. Compared to this, array of input values represent array of state variables. Each discrete state variable is sampled with predefined step form the interval $\langle 0, 1 \rangle$.

Operation in non-Episodical Experiments: The Q-Learning belongs into the group of algorithms which learn episodically. At the beginning of each episode, the SRP should start to operate from randomly chosen state of the environment. This ensures that the algorithm learns efficiently in the entire state-space. However, in real-life experiments this cannot be provided often. The successful learning in continuous experiments is accomplished by connecting to the own motivation source. By dynamically adjusting the *exploration vs. exploitation* tradeoff, the learning can be efficient.

In a particular architecture, the RL module represents some behaviour. We introduce the parameter called **Importance**. Increasing of this parameter affects two following components in the Neural Module:

- Causes **decrease of ϵ parameter** in the ϵ -Greedy ASM. Therefore, when the behaviour represented by the module has high importance, the exploration is suppressed.
- Causes **increase of value of the selected action**. This ensures that in competition against other RL modules (or other action-selecting sub-systems) has higher chance to win.

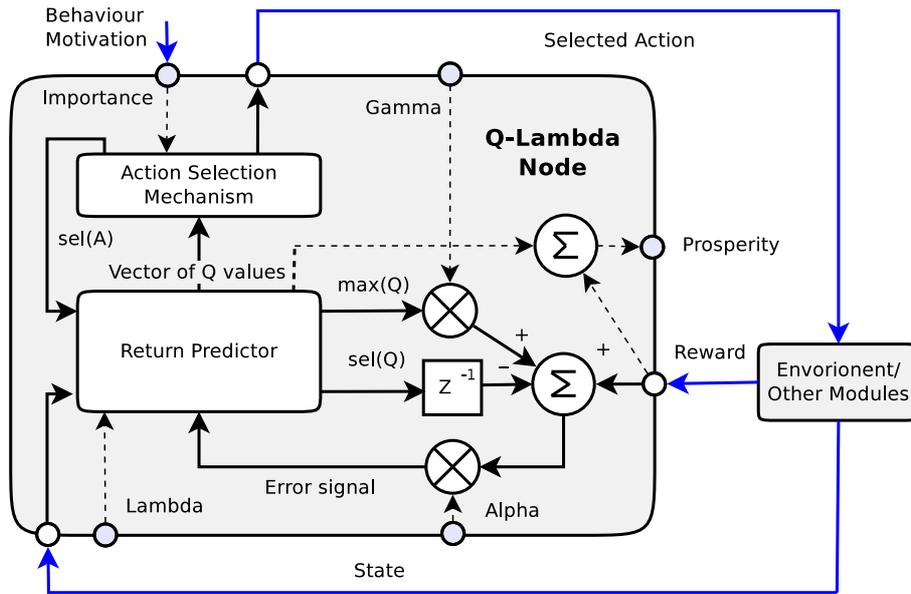


Fig. 2. Scheme of the Q-Learning system. The line labeled “max(Q)” is the prediction of return for the best action. The “Sel(Q)” is the Q actually taken, it is combined with return prediction and reinforcement received from the environment r_i through unit the delay z^{-1} , γ is the discount factor and α is the learning rage. The predictor predicts action values in a current state, based on this information the ASM selects action to be executed. The corresponding Neural Module will have the following configuration inputs: α, γ, λ , data inputs will represent states and reward and data outputs will represent actions to be executed.

Defining Prosperity of Q-Learning Neural Module:

The single value of Prosperity for this module can be difficult to choose. In our implementation, the prosperity of the module is composed of two values as follows:

$$P_t = \frac{Cover_t + MCR_t}{2}, \quad (5)$$

Where, Mean Cumulative Reward (MCR) is defined as mean reward (R) received during the simulation:

$$MCR_t = \frac{\sum_i R_i}{i} \quad \forall i \in 0..t, \quad (6)$$

and the $Cover_t$ represents how many states has been visited (by the RL module) during the simulation.

Motivation Source Module Design

The only change in the motivation source implementation (compared to the theory) is that the Module produces two data outputs: **The motivation** for the behaviour and **the reward** received on its input. The reward output serves mainly for simpler connecting of modules in the simulator.

The prosperity of this module should correspond to the value of physiological variable V , defined in the equation 3. The *limbo* represents the optimal conditions of agents physiology. If the physiological state space is in the limbo area, no motivation is produced Kadleček (2008). Here the limbo area is in $V = 1$. The *Mean State Distance* (MSD) to optimal conditions (limbo area) is defined as:

$$MSD_t = \frac{\sum_i SD_i}{i} \quad \forall i \in 0..t, \quad (7)$$

where SD_i is state distance from the optimal conditions.

$$P_t = 1 - MSD_t. \quad (8)$$

The more time spent near the optimal conditions, the better agents behaviour probably is. Therefore the Prosperity of the Motivation source is defined in the equation (8).

EXPERIMENTS

The resulting modules were tested in an architecture composed of one Module with motivation source and one Q-lambda module. The importance input and the reward input of the Q-lambda module was connected to the Motivation source module. Connecting the motivation to the importance input causes that the **action selection is dynamically weighted between the greedy and randomized** one.

Experiment Description

The architecture was tested on discrete grid map of size 20×20 with obstacles and one attractor, the environment is described in the Fig.3. The agent was equipped with 4 actions (moving in four directions) and the reward was received after reaching the position containing the reward.

Concluded simulations are made as non-episodical, this means that the agent starts on initial position (center of the map) and is let to interact with the environment freely for given number of steps.

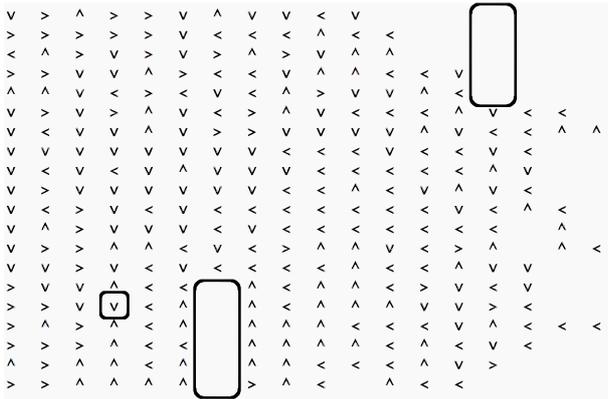


Fig. 3. Simple 2D environment representation. Agent has ability to execute four actions (moving in four directions). Each position in the map contains symbol representing action with the highest $Q(s, a)$ value learned. When the agent follows the greedy policy (Greedy ASM), these actions will be used. In the map, there are two obstacles and one attractor.

Validating Functionality of Learning Algorithm

The presented values of prosperity are presented from 10 non-episodical experiments, each started from the same initial state (center of the map) and lasted 100000 discrete steps. The RL algorithm was configured with the following empirically-estimated constant parameters: $\alpha = 0.5$, $\gamma = 0.3$, $\lambda = 0.04$.

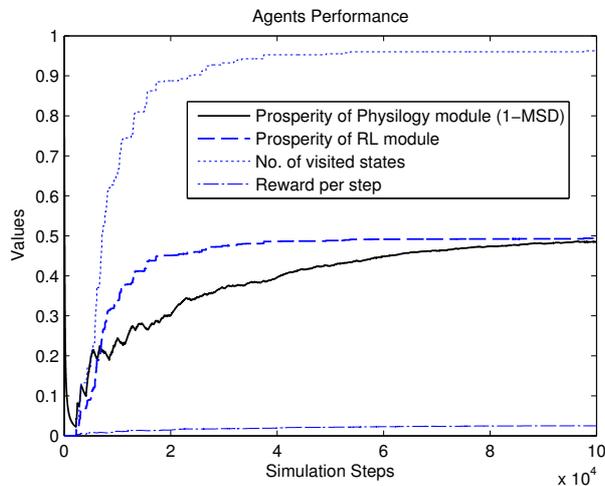


Fig. 4. Typical course of Prosperities of particular Neural Modules in the architecture. The complete Prosperity value is composed of prosperities of the Q-Lambda Module and the Physiology Module. The value should represent how suitable is overall architecture design for a given task. In this case, the value also should represent the convergence of learning. We can see that despite the value of Reward/step is small, the agent learns to maintain its physiology in reasonable distance from optimal conditions. It can be also seen that the agent explores the environment efficiently and it has visited almost all states (except obstacles).

The prosperity of architecture is defined as a sum of prosperities of all Neural Modules used:

$$P_{architecture} = \frac{\sum_i P_i}{i} \quad \forall i \in architecture, \quad (9)$$

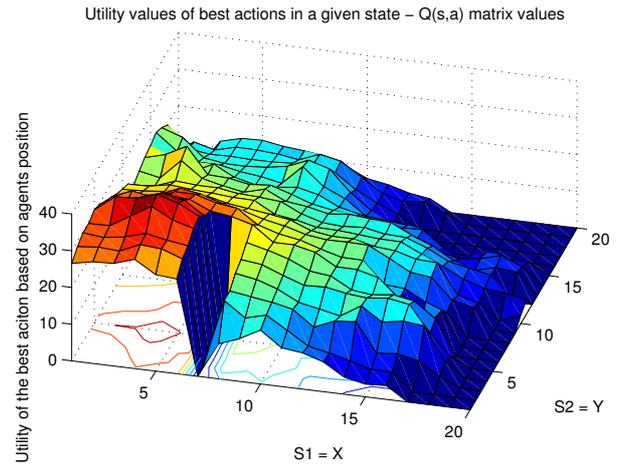


Fig. 5. Value of the highest utility (of the best action a^*) in the $Q(s, a)$ matrix, based on agents X, Y position in the map. The nearer to the reward source, the higher the expected outcome of the best action is. Positions with obstacles have value of $Utility = 0$. Note that the Greedy ASM selects the action with the highest utility in each state. For better visibility, values on the Z axis are rescaled from the interval $(0, 1)$. These values represent utilities of actions depicted in the Fig.3.

where P_i is the prosperity of one module. Here, the Q-Lambda and the Motivation source Modules were used. The value should represent how suitable is overall architecture design for a given task. In this case, the Prosperity should also represent the convergence of agents learning.

Note that the Prosperity of Q-Lambda module is composed of MCR and $Coverage$ values. Therefore its value represents how often the module (agent) receives the reward as well as how many states it has covered. The value of the Motivation source is the MSD , which represents how "satisfied" is the agent with its behavior.

The Fig.3 is a graphical representation of the best (learned) action in the state (X, Y) coordinates). This represents agents autonomously learned knowledge during 10000 simulation steps, this strategy would be followed in case that the Greedy ASM was used. The Fig.5 depicts values in the $Q(s, a)$ matrix. The highest Utility value for each state (X, Y) coordinates) is depicted, these values correspond to the actions shown in the Fig.3.

Validating the Purpose of Motivation

During the simulation, the amount of motivation determines how important is learned behaviour. After obtaining the reward, the motivation decreases towards zero and the agent starts to perform the exploration. The value of the motivation should directly correlate to the agents policy. In order to test this, the speed of decay of the physiological variable (see the equation (3)) was altered and resulting agents behaviour was observed.

In case when the decay of the variable is fast, the agent should "switch" to the exploitation of learned knowledge frequently. After satisfying the motivation, the agent returns to the exploration.

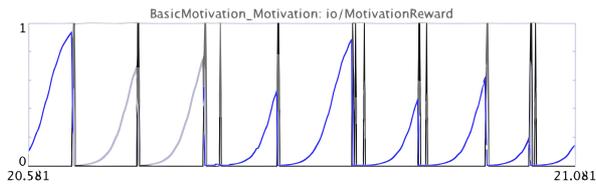


Fig. 6. Motivation value in case when the motivation increases fast. The X axis represents time steps and the Y represents value of particular variable. Spikes in the graph represent binary event of receiving the reward. Continuous value (sigmoidal curves) represents the amount of motivation. It can be seen that the motivation is source of exploitation of behaviour which leads to the reward.

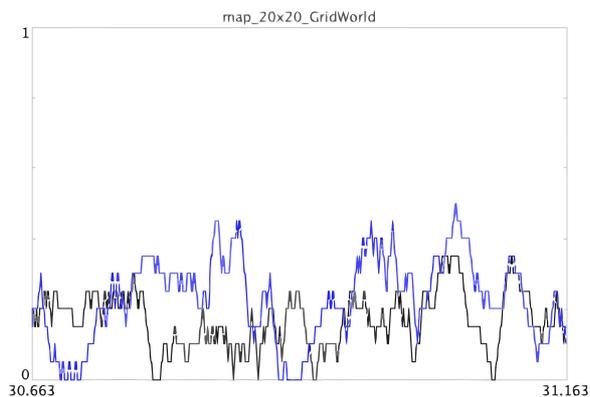


Fig. 7. Agents position in the map in time, which corresponds to the fast increasing motivation. The X axis represents time steps and the Y axis represents agents current position. Two lines represent the X and Y position in the map. Note that values of state variables are sampled from interval of (0, 1) and the source of reward is located on coordinates (3, 4). It can be seen that the agent tends to stay near the reward source most of the time.

The Fig.6 shows the case when the $decay = 0.01$ causes relatively fast increase of motivation. When the motivation is low, the agent explores the environment. As the motivation increases, gradually less randomization is used until the agent reaches the reward. This sets the motivation back towards zero. The Fig.7 roughly corresponds to the Fig.6 and represents the agents position in map during the simulation. It can be seen that the agent explores only states that are near the motivation source.

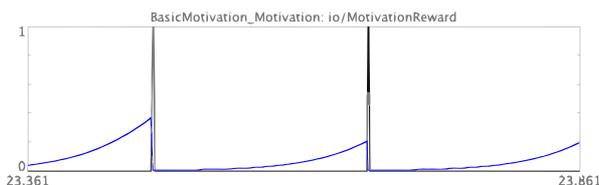


Fig. 8. Motivation value increases slowly in this case, therefore the agent has "more time" for exploration. The physiology is satisfied (by obtaining the reward) only when needed.

In the following experiment, the value of $decay =$

0.002 causes slower increase of motivation. The Figures 8 and 9 show agents behaviour in this case. It can be seen that the agent has less overall motivation to obtain the reward and exploitation of the behaviour is less important. The Fig.9 shows that the agent explored bigger part of the environment while satisfying the motivation when necessary.

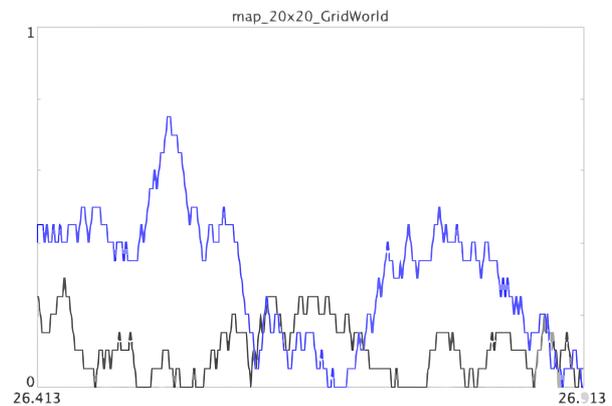


Fig. 9. Agents position in the map in time. Compared to the Fig.7, here the motivation increases slower. In this case, the agent has more time to explore and therefore reaches more distant parts of the environment too.

CONCLUSIONS

Here, the two reusable Neural Modules were presented, one implementing the Q-Learning algorithm and the second implementing the source of motivation. The functionality of these modules was experimentally tested. The results show that the agent "test-architecture" composed of these modules is able to successfully learn in discrete non-episodic experiments (Fig.3,4,5). The expected oscillations between exploration and exploitation behaviour were observed (Fig.6,8). This implies that the architecture was able to dynamically switch between the knowledge exploration and exploitation as needed (Fig.7,9).

These presented modules are compatible with our HANNS framework, but can be also used as stand-alone ROS nodes. The Java implementation of these nodes is freely available online (together with the environment) at <https://github.com/jvitku/rl>, and the Motivation source available at: <https://github.com/jvitku/physiology>.

The presented Neural Modules are made as domain-independent as possible and therefore may be directly incorporated in new architectures of autonomous agents, such as those proposed in Kadleček (2008) or Vítků (2011). Furthermore, the optimization techniques can be applied in order to build new architectures for a particular task. By means of techniques similar to the EAs, the modules can be used in searching for entirely new agent architectures.

ACKNOWLEDGEMENT

This research has been funded by the Dept. of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, under the SGS Project SGS14/144/OHK3/2T/13.

REFERENCES

- Auda, G. and Kamel, M. (1999). Modular neural networks: a survey. *Int J Neural Syst*, 9(2):129–151.
- Deb, K. (2011). Multi-objective optimisation using evolutionary algorithms: An introduction. In Wang, L., Ng, A. H. C., and Deb, K., editors, *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*, pages 3–34. Springer London.
- Eliasmith, C. and Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. The MIT press, Cambridge, ISBN: 0-262-05071-4.
- Fekiac, J., Zelinka, I., and Burguillo, J. C. (2011). A review of methods for encoding neural network topologies in evolutionary computation. In *Proceedings 25th European Conference on Modelling and Simulation ECMS*, pages 410–416. ISBN: 978-0-9564944-2-9.
- Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions of Neural Networks*, 14:1569–1572.
- Kadleček, D. and Nahodil, P. (2008). Adopting animal concepts in hierarchical reinforcement learning and control of intelligent agents. In *Proc. 2nd IEEE RAS & EMBS Int. Conf. Biomedical Robotics and Biomechatronics BioRob 2008*, pages 924–929.
- Kadleček, D. (2008). *Motivation driven reinforcement learning and automatic creation of behavior hierarchies*. PhD thesis, Czech Technical University in Prague, FEE. Supervised by Pavel Nahodil.
- Kadleček, D. and Nahodil, P. (2001). New hybrid architecture in artificial life simulations. In *Advances in Artificial Life. In Lecture Notes in Artificial Intelligence*, volume 2159, pages 143–146. Berlin: Springer Verlag, Berlin, Germany. ISBN: 978-3-540-42567-0.
- Kordík, P. (2006). *Fully automated knowledge extraction using group of adaptive models evolution*. PhD thesis, Czech Technical University in Prague, FEE, Dep. of Comp. Sci. and Computers.
- Maass, W. (1996). Networks of spiking neurons: The third generation of neural network models. In *Journal Neural Networks*, 10:1659–1671.
- Murre, J. M. J., Phaf, R. H., and Wolters, G. (1989). Calm networks: a modular approach to supervised and unsupervised learning. In *Proc. Int Neural Networks IJCNN. Joint Conf*, pages 649–656.
- Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y. (2009). Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.
- Thomas, D. B. and Luk, W. (2009). Fpga accelerated simulation of biologically plausible spiking neural networks. In *In Proc. IEEE Symp. Field-Programmable Custom Computing Machines (FCCM)*.
- Vítků, J. (2011). An artificial creature capable of learning from experience in order to fulfill more complex tasks. Diploma thesis, Czech Technical University in Prague, FEE, Dept. of Cybernetics. Supervised by Pavel Nahodil.

AUTHOR BIOGRAPHIES

JAROSLAV VÍTKŮ was born in Prague, Czech Republic. He graduated in 2011 in Czech Technical University in Prague, Faculty of Electrical Engineering in Artificial Intelligence. His diploma thesis "An Artificial Creature Capable of Learning from Experience in Order to Fulfill More Complex Tasks" was awarded by Price of Dean. Currently, he is a PhD student at the Department of Cybernetics, CTU in Prague, still under the guidance of his supervisor Pavel Nahodil. Here elaborates the results of his thesis as an automated design of complex modular systems inspired by Nature. His research interest includes hybrid neural networks, cognitive science, biologically inspired algorithms, behavioral robotics and Artificial Life in common. Now, he is finishing his dissertation: "Autonomous Design of Modular Agent Architectures". Some interesting parts of the thesis are presented in this paper for the first time. His e-mail address is: vitkujar@fel.cvut.cz.

PAVEL NAHODIL was born in Prague, Czech Republic. Since 1970 he has been working at the Department of Cybernetics at the Faculty of Electrical Engineering, Czech Technical University in Prague, where he was also appointed the Professor in Technical Cybernetics (in 1986). He has led and consulted more than 123 diploma thesis here so far. He was also supervisor of about 30 PhD doctoral students till now. His present professional interest includes artificial intelligence, multi-agent systems, on behavior based intelligence robotics (control systems of artificial creatures) and the artificial life design in general. He is (co-)author of several books, university lecture notes, hundreds of scientific papers and large collection of scientific studies. He is also the organizer of some international conferences + reviewer (IPC Member) and a member of many Editorial Boards. His e-mail address is: nahodil@fel.cvut.cz.

ROUTING AND COMMUNICATION PATH MAPPING IN VANETS

Nnamdi Anyameluhor and Dr Evtim Peytchev
School of Science and Technology
Nottingham Trent University
Clifton Lane, Nottingham, NG11 8NS, UK
E-mail: {nnamdi.anyameluhor2012, evtim.peytchev}@ntu.ac.uk

KEYWORDS

Vehicular Ad-Hoc Network; Cross-layer; Routing; Network Simulation; ITS; NS-3; Communication Mapping; MAC; Mobile Programming;

ABSTRACT

Vehicular ad-hoc network (VANET) has quickly become an important aspect of the intelligent transport system (ITS), which is a combination of information technology, and transport works to improve efficiency and safety through data gathering and dissemination. However, transmitting data over an ad-hoc network comes with several issues such as broadcast storms, hidden terminal problems and unreliability; these greatly reduce the efficiency of the network and hence the purpose for which it was developed. We therefore propose a system of utilising information gathered externally from the node or through the various layers of the network into the access layer of the ETSI communication stack for routing to improve the overall efficiency of data delivery, reduce hidden terminals and increase reliability. We divide route into segments and design a set of metric system to select a controlling node as well as procedure for data transfer. Furthermore we propose a system for faster data delivery based on priority of data and density of nodes from route information while developing a map to show the communication situation of an area. These metrics and algorithms will be simulated in further research using the NS-3 environment to demonstrate the effectiveness.

INTRODUCTION

When the Fleetnet project in Germany promised better route information, traffic and road condition updates and socialising on the road in 2001 (Hartenstein et al. 2001) it seemed a bit futuristic but nowadays we see that these kinds of vehicular services should be a necessity in order to increase comfort and safety of road users as well as greatly improve traffic management systems. This 'lofty' idea is part of the Intelligent Transport System that has been developed and is continually evolving as a lot of research and interest is shown in the area. The Intelligent Transport System (ITS) is simply the combination of telecommunication systems, information technology and transport system to

provide relevant information to road users for the sake of safety and efficiency (Bishop 2005).

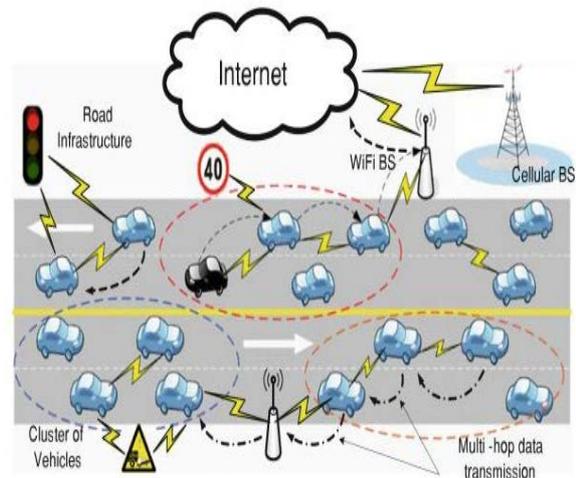


Figure 1: VANET Concept (Zhu and Li 2013)

In order to disseminate information two methods can be used which are cellular networks and other wireless short-range definitions such as those defined by the Cooperative Intelligent Transport Systems (C-ITS). Whereas cellular networks work through a centralised topology with standard infrastructure (Base Station etc.) regardless of the position of the transmitting and receiving nodes, C-ITS on the other hand defines a framework where direct wireless networked devices can form an ad-hoc networks. When vehicles form ad-hoc networks by connecting to each other within range it is called a vehicular ad-hoc network (VANET), it is a highly mobile network where links can be created and broken in seconds (Karagiannis et al. 2011) therefore it is of high importance to have a system that can cope with this mobility while being able to maintain connection and data flow. In 2009 the European commission issued a formal request (Mandate 453) to CEN and ETSI, to prepare a set of standards and specifications to support the implementation of Cooperative ITS systems across Europe (ETSI 2014) The ETSI have developed an adapted OSI network framework to accommodate ITS applications and design as seen in figure 2

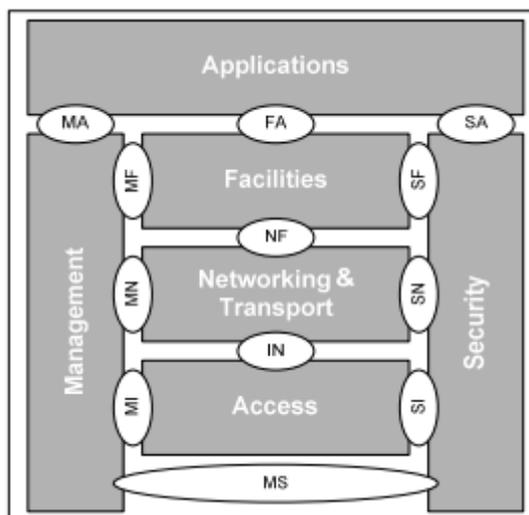


Figure 2: ITS Architecture

Data loss, packet collision, broadcast storms and hidden terminal problems are some major issues that may be faced in vehicular ad hoc networks because the nature of the wireless medium as well as change in network topology due to the obvious reasons of mobility and speed (Yousefi et al. 2006). Pure flooding broadcast and other traditional broadcast schemes (without acknowledgment) may not be efficient enough to ensure reliable data spread (Karagiannis et al. 2011). Also the communication knowledge of an area in VANETs has been neglected in research, this knowledge is very important in terms of message dissemination and other decision-making processes.

In this paper we propose a system for utilisation of cross layer information from the physical layer and network layer to improve cluster formation and for relaying of data; this also includes data priority while service messages are in a contention loop as well as channel selection process. We also propose the use of node density algorithms to construct a communication map of the area.

RELATED WORK

Efficiency and reliability are important factors in the design of VANET protocols, especially in cases of emergencies. Unfortunately these factors are not adequately catered for in the default architecture due to the use of a MAC protocol based on IEEE 1609.4 and IEEE 802.11p (Sjöberg 2013). This protocol employs the use of CSMA in the MAC layer which has been shown to perform poorly as the number of nodes increase in the network due to an equally increased probability of two or more nodes having the same back-off counter for attempted packet delivery (Bilstrup et al. 2008), the result of which is packet collisions.

This disadvantage has led to a lot of research focused on improving the throughput of packets in the network.

Among these research Zhou et al. (2008) proposed a cross layer cooperative VANET with rate and medium access control, they introduced cooperative communication and cross layer link detection at the MAC, network and transport level in order to improve message dissemination and fairness of medium access. However, this system employs the use of roadside units or “gateways” to achieve this aim, this will no doubt add to the overall cost of the VANET architecture and reliance on extra hardware should as much as possible be minimised. The AMB protocol (Korkmaz 2009) requires a node to propagate a “black burst” or jamming signal as far as possible within the network with the furthest node responding and hence becoming the next relay for messages. The drawback of this system is the time delay involved in the sending and response of this black burst. In research done by Bi et al. (2009) nodes have reduced time delay but the work focuses message propagation based on direction. De Couto et al. (2005) proposed metric for high throughput in multi-hop networks by taking into account the probability ratio of expected transmission count in effective message delivery to the number of attempted transmission. Cross layer broadcast protocol by Bi et al. (2010) describes a system for efficient routing and network organisation in order to tackle the hidden terminal problem and the broadcast storm issue by using a request to send and clear to send (RTS/CTS) scheme. This RTS/CTS is their response to the lack of acknowledgement in the CSMA protocol. Our protocol combines cross layer functionality to provide efficient and reliable routing; cluster aggregation of nodes to improve fairness, avoid broadcast storms and hidden terminals; TDMA and transmission request priority to avoid collision and finally a system to generate communication maps of an area.

PROPOSED WORK

Cluster head Selection

Hafeez et al. (2011) proposed a cluster and OFDMA MAC protocol with each cluster assigned a set of subcarrier to distinguish it from other clusters, cluster heads are established based mainly on a weighted stability factor on the road. While the system increases reliability and reduces delay there is the issue of the system adapting or performing self-learning processes to maintain this reliability. Also due to wireless ranges the choice of making each cluster the size of two times the range of the cluster head's transmission may be inhibiting as the cluster head may struggle to reach every member of the cluster.

For this research we also intend to create a cluster of nodes by dividing the network into small groups with a viable cluster head whose primary duty includes maintaining stability of the network and data flow. By creating small clusters we hope to form and maintain a

network within a short period while enabling a more robust system for data dissemination along the route.

In Figure 2 it can be noted that there is a Facility layer in the ETSI model and among the duties of the Facilities layer include; support of getting and combining data, channel selection etc. Therefore the following description can be processed and shared within the facility layers of nodes in a one hop radius with respect to selecting a cluster head.

Since the ETSI standard covers an application sensitive layer, we propose an internal control worthiness metric to calculate which node should be the cluster head within each cluster. The data used in this calculation will include Nodal Age on Route (NAR), Position Relative to other Nodes (PRON), Direction (DIR), Speed and Acceleration (SACC) and Node Equipment Status (NES); we assume the presence of these information and will set them dynamically for this research. Algorithms and processes of how to gather data have been done in research carried out by Gamati (2012). Together the information is used to develop a Master Probability Metric (MPM) system by assigning values and corresponding Impact Factors (IMF) to the information and sorting them according to the nodes with the highest impact, the highest then selected as the cluster head.

Table 1: Information, Value and Impact Factor for Determining Suitable Cluster Head.

INFORMATION	VALUE	IMPACT FACTOR
Position	Central	9
	Corner	1
Direction	Corresponding	9
	Opposite	1
Speed and acceleration relative to other nodes	High	1
	Steady	9
Age on route (communication map metric)	Lengthy	9
	Short	1

From Table 1 it can be noted that a central node (relative to other nodes) with a steady speed and low acceleration has a better chance of being elected a cluster head. In order to calculate these process we will consider the selection as a multi-dimensional knapsack issue where the aim is to maximise a choice outcome based on several constraints:

Maximise

$$\sum_{j=1}^n P_j X_j \quad (1)$$

Subject to:

$$\sum_{j=1}^n W_{ij} X_j \leq W_i, \text{ for all } 1 \leq i \leq m \quad (2)$$

$$X_j \geq 0, X_j \text{ integer for all } 1 \leq j \leq n$$

The algorithm takes the summation of the values and impact factors and must pick a solution while satisfying the stated constraints. Given that this computation may take up to a few seconds to complete, this system may be insufficient for emergency purposes and better suited to nodes in a platoon formation.

Data Routing

According to the ETSI standard for C-ITS message priority may be handled in two ways which are; transmission request from the Data Link Layer and those from the PHY layer both in the Access block. The transmission request from the DLL handles messages from the 'Application' block of the station while the request from the PHY layer handles contention of different ITS stations in the physical communication channel (ETSI 2013). This second transmission contention between different nodes is what we look into in this proposal.

At the end of the cluster head selection process and network stabilization, data may now be routed among members of the cluster and surrounding with TDMA through the cluster head using the CALM (Communication Access for Land Mobiles) M5 protocol. We propose the use of TDMA in place of CDMA in order to avoid unnecessary packet collisions in the system. In this process every message is given a class and priority value e.g. 1 for emergency, 2 for critical messages, 3 for network initialisation etc. this is done for the sole purpose of facilitating routing at the MAC level and making sure priority messages are given due resources.

Table 2 Example of Data Type and Priority Value

Data Type	Priority value
Hello packets	1
High node impact	2
High braking	3
.	.
.	.
.	.
.	.
Proprietary messages	n

The messages are polled by the cluster head and shared in the cluster according to the process described by the flow chart in Figure 3.

The highest priority message is given the next available time slot and resource while other data is reorganised accordingly with the exit or addition of packets.

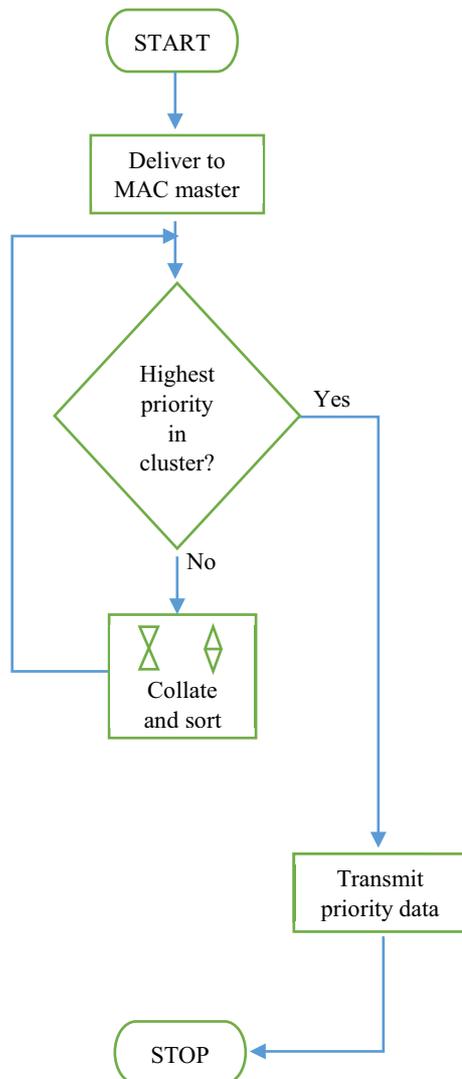


Figure 3: Flow chart for data transmission

COMMUNICATION MAPPING

Mapping has always been a simple way of communication information to an end user and the focus of work so far has been on providing useful information and traffic condition information, however we believe that the nodes can provide other useful information to aid building of networks across larger areas, hence in this case we propose a system to draw up not routing

information but communication information for a section of a city based on both prior and current information example seen in the Figure 4.

By using node density algorithms and past node history we aim to be able to draw a map to describe how the communication network in a certain area is at a particular time e.g. Assuming each coloured path represents a route within a city, this paths have had their communication situation analysed with respect to node density, gathered node information, channel conditions etc. we represent the conditions as particular colours in this case red represents very poor communication conditions and green represents good communication condition.

This we hope will help reduce the decision time for routing and network management

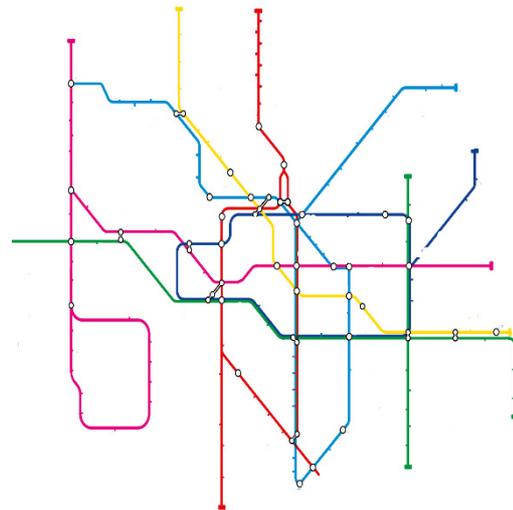


Figure 4: Communication map model with different colours indicating different network conditions

CONCLUSION

In this paper we have proposed a system to create and maintain network and routing stability for nodes in a platoon arrangement by utilising data from the physical world; it involves selecting a Mac controller or cluster head and forming node segments. We also proposed the use of TDMA for the control and service channels in VANETs in place of the original CDMA due packet collisions and reliability. Finally we have also proposed a communication mapping system based on node density algorithms. Our future work will be to simulate these proposals using NS3 and measure the performance against both the IEEE 802.11p standard and other popular suggestions. The concept of VANET is one that aims to reduce accidents, aid better traffic management and on-road sales and services, therefore will benefit from a more stable design as proposed.

REFERENCES

- Agarwal, A., Starobinski, D. and Little, T.D.C. (2007) 'Exploiting downstream mobility to achieve fast upstream message propagation in vehicular ad hoc networks', in 2007 Mobile Networking for Vehicular Environments, May, 11 November. Anchorage, AK, pp.13–18.
- Bishop, R. (2005) *Intelligent Vehicle Technologies and Trends*, Artech House, Boston, Mass, USA.
- Bi, Y., Cai, L.X., Shen, X. and Zhao, H., 2010. A cross layer broadcast protocol for multihop emergency message dissemination in inter-vehicle communication. In: *Communications (ICC), 2010 IEEE International Conference on, IEEE*, pp. 1-5.
- Bi, Y., Zhao, H. and Shen, X., 2009. A directional broadcast protocol for emergency message exchange in inter-vehicle communications. In: *Communications, 2009. ICC'09. IEEE International Conference on, IEEE*, pp. 1-5.
- Bilstrup, K., Uhlemann, E., Strom, E.G. and Bilstrup, U., 2008. Evaluation of the IEEE 802.11 p MAC method for vehicle-to-vehicle communication. In: *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th, IEEE*, pp. 1-5.
- De Couto, D.S., Aguayo, D., Bicket, J. and Morris, R., 2005. A high-throughput path metric for multi-hop wireless routing. *Wireless Networks*, 11 (4), 419-434.
- ETSI, 2013. *Intelligent Transport Systems (ITS); Communications Architecture Standard* [online]. ETSI. Available at: http://www.etsi.org/deliver/etsi_en/302600_302699/302665/01.01.01_60/en_302665v010101p.pdf [Accessed Feb/4 2014].
- ETSI Cooperative ITS 2014. [Online]. ETSI. Available at: <http://www.etsi.org/technologies-clusters/technologies/intelligent-transport/cooperative-its> [Accessed February/4 2014].
- Gamati, E., Peytchev, E. and Germon, R., 2013. Utilization of Broadcast Methods for detection of the road conditions in VANET. Nottingham Trent University - School of Science and Technology - Computing and Informatics.
- Hafeez, K.A., Zhao, L., Liao, Z. and Ma, B.N., 2011. Clustering and OFDMA-based MAC protocol (COMAC) for vehicular ad hoc networks. *EURASIP Journal on Wireless Communications and Networking*, 2011 (1), 1-16.
- Hartenstein, H., Bochow, B., Ebner, A., Lott, M., Radimirsch, M. and Vollmer, D., 2001. Position-aware ad hoc wireless networks for inter-vehicle communications: the Fleetnet project. In: *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing, ACM*, pp. 259-262.
- Karagiannis, G., Altintas, O., Ekici, E., Heijenk, G., Jarupan, B., Lin, K. and Weil, T., 2011. Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions. *Communications Surveys & Tutorials, IEEE*, 13 (4), 584-616.
- Korkmaz, G., Ekici, E. and Ozguner, F., 2007. Black-burst-based multihop broadcast protocols for vehicular networks. *Vehicular Technology, IEEE Transactions on*, 56 (5), 3159-3167.
- Sjöberg, K., 2013. *Medium Access Control for Vehicular Ad Hoc Networks*. Doctor of Philosophy, Chalmers University of Technology; Halmstad University.
- Taniguchi, E., Thompson, R.G., Yamada, T. and Van Duin, R., 2001. *City Logistics. Network modelling and intelligent transport systems*.
- Yousefi, S., Mousavi, M.S. and Fathy, M., 2006. Vehicular ad hoc networks (VANETs): challenges and perspectives. In: *ITS Telecommunications Proceedings, 2006 6th International Conference on, IEEE*, pp. 761-766.
- Zhou, L., Zheng, B., Geller, B., Wei, A., Xu, S. and Li, Y., 2008. Cross-layer rate control, medium access control and routing design in cooperative VANET. *Computer Communications*, 31 (12), 2870-2882.
- Zhu, H., and Li, M., 2013. *Studies on Urban Vehicular Ad-hoc Networks*. 1st Ed. New York: Springer New York.

AUTHOR BIOGRAPHIES



DR. EVTIM PEYTCHEV is Reader in Wireless, Mobile and Pervasive Computing in the school of Science and Technology at Nottingham Trent University, UK. He is leading the Intelligent Simulation, Modelling and Networking Research Group, which consist of 5 lecturers, 3 Research Fellows and 6 research students. He is the Module Leader for Systems Software; and Wireless and Mobile Communications. He also teaches on the modules Software Design and Implementation; Mobile Networking; Enterprise Computing; and Computer Architecture.



NNAMDI ANYAMELUHOR went to the University of Science and Technology (Nigeria) where he received his B.Tech in Electrical and Electronic Engineering, he then went to study Cybernetics and Communication Engineering at Nottingham Trent University (United Kingdom) where he obtained his M.Sc. He is currently a Post Graduate researcher in Nottingham Trent University in the field of Vehicular Ad Hoc Networks.

MODELLING RETINAL FEATURE DETECTION WITH DEEP BELIEF NETWORKS IN A SIMULATED ENVIRONMENT

Diana Turcsany and Andrzej Bargiela
School of Computer Science
The University of Nottingham
NG81BB, Nottingham, United Kingdom
Email: {dxt, abb}@cs.nott.ac.uk

Tomas Maul
School of Computer Science
The University of Nottingham Malaysia Campus
43500, Semenyih, Malaysia
Email: Tomas.Maul@nottingham.edu.my

KEYWORDS

Retina model; Neural learning; Deep belief network; Probabilistic modelling; Simulated environment

ABSTRACT

Recent research has demonstrated the great capability of deep belief networks for solving a variety of visual recognition tasks. However, primary focus has been on modelling higher level visual features and later stages of visual processing found in the brain. Lower level processes such as those found in the retina have gone ignored. In this paper, we address this issue and demonstrate how the retina's inherent multi-layered structure lends itself naturally for modelling with deep networks. We introduce a method for simulating the retinal photoreceptor input and show the efficacy of deep networks in learning feature detectors similar to retinal ganglion cells. We thereby demonstrate the potential of deep belief networks for modelling the earliest stages of visual processing.

INTRODUCTION

Vision systems of humans and other biological systems have long been within the centre of interest in neuroscience and biology. The main focus of this work is the first stage of visual information processing in mammals, implemented by the retina.

Extending our knowledge of visual information processing mechanisms within biological systems has key importance in finding suitable learning algorithms for artificial retinæ. Such algorithms are much sought after, as building better computational models of the retina is not only crucial for advancing the field of retinal implant engineering, but could also improve image processing and computer vision algorithms.

Despite discoveries regarding the anatomy and physiology of neural structures, understanding biological vision systems remains an open problem. Due to the high number, diverse functionality and complex connection patterns of neurons, our knowledge regarding the roles of cells and circuits of the visual pathway is still incomplete. Even well studied processing units, such as the retina are not fully understood and recent studies have revealed a lot more complexity in the functionality implemented by retinal cells than traditionally believed (Gollisch and Meister, 2010).

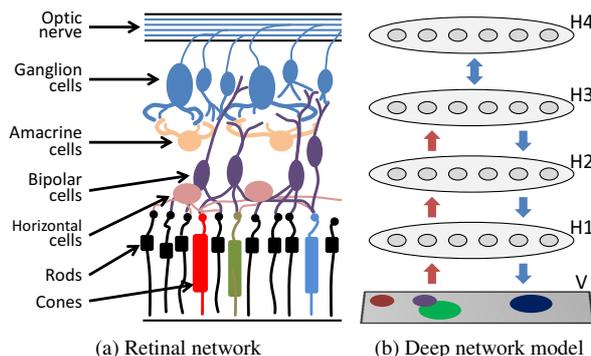


Figure 1: (a) Schematic of interacting cell types in the multi-layer retinal network. (b) Diagram showing one example of our retina models: a 4 hidden layer deep belief network that generates images of our dataset together with their class labels. V denotes the visible and H_1-H_4 the hidden layers. Blue arrows correspond to the generative model, while the upward pointing red arrows indicate the direction of inference (class label prediction). In the generative model the top two layers form an associative memory.

As a consequence of this incomplete knowledge, currently the design of computational retina models involves dealing with a great amount of uncertainty and only partially known information. In such an environment our aim is to develop computational models which exhibit substantial fidelity to the retina's currently known neural structure, while also being highly adaptable. We anticipate that flexible probabilistic models, such as deep belief networks (Hinton et al., 2006) studied here, can provide the required adaptive power.

Modelling The Retinal Network

Mammalian retinæ contain 6 main cell types: rods, cones, and horizontal, bipolar, amacrine, and ganglion cells, which are organised into consecutive layers as illustrated in figure 1a. The first layer contains photoreceptors: rods and cones. Rods are essential for sensing light in darker environments, whereas cones provide the first processing step in colour vision. The retinæ of humans and trichromatic monkeys contain 3 types of cones: blue cones sense short, green cones medium and red cones long wavelength light. Colour vision is possible by comparing the response of different wavelength cones.

The synaptic connections between the photoreceptor layer and the horizontal and bipolar cells form the outer plexiform layer. The main role of horizontal cells is in local gain control, which guarantees the input signal reaching the further processing units is kept within the appropriate range. Bipolar cells receive input from photoreceptors and in the inner plexiform layer they connect onto ganglion and amacrine cells. Bipolar cells have centre-surround receptive field organisation and can either connect to the ON or the OFF pathway, which are respectively responsible for detecting light objects on dark background and vice versa.

Amacrine cells serve important roles in various tasks, e.g. object motion detection (Ölveczky et al., 2003). The highest level of information processing within the retina is implemented by ganglion cells. These cells receive input from a number of bipolar cells and occasionally from amacrine cells. Different ganglion cell types detect specific visual features and transmit the extracted information through the optic nerve towards higher processing areas located in the visual cortex. The most well-known ganglion cells are the ON/OFF local edge detectors, exhibiting centre-surround antagonism. ON-centre cells receive excitatory signals when light appears in their receptive field centre and inhibitory signals resulting from light in the surround, while the opposite is true for OFF-centre ganglion cells. The response patterns of these cells are most often modelled as Difference-of-Gaussians (DoG) filters (see figure 2).

For more in depth descriptions of the retinal network, please refer to review papers by Kolb (2003), Wässle (2004) and Masland (2012).

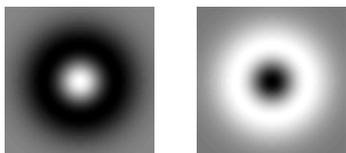


Figure 2: Difference-of-Gaussians (DoG) filters: common models of ON (left) and OFF (right) type ganglion cells.

Neural computation units in the retina have been modelled at different abstraction levels, ranging from detailed models of single neurons (Kameneva et al., 2011; Usui et al., 1996; Velte and Miller, 1997), to populations of neurons (Pillow et al., 2008; Shlens et al., 2008) and networks of interacting retinal cell types (Gaudiano, 1992; Kien et al., 2012; Maul et al., 2011; Teeters et al., 1997). Our work focuses on higher level description of retinal circuits and assesses large-scale network models.

Multi-layer Networks

To learn a multiple layer generative model of the data where each new layer corresponds to a higher level representation of information, Hinton et al. (2006) train a deep belief network (DBN) by “greedy” layer-by-layer learning using the unsupervised restricted Boltzmann machine

(RBM) algorithm. The network parameters obtained from the unsupervised learning phase are subsequently fine-tuned using backpropagation. The potential of DBNs for learning meaningful features is demonstrated on visual recognition tasks and text retrieval (Hinton and Salakhutdinov, 2006). Since Hinton et al. (2006)’s efficient training method for deep networks was introduced, there has been increasing research in this direction. Successful applications of deep belief networks and other deep architectures (Lee et al., 2009a; Salakhutdinov and Hinton, 2009) have been presented on visual tasks (Eslami et al., 2012; Kavukcuoglu et al., 2010; Krizhevsky et al., 2012; Le et al., 2012) and on a number of machine learning problems (Collobert and Weston, 2008; Larochelle et al., 2007; Lee et al., 2009b; Salakhutdinov and Hinton, 2007).

A MULTI-LAYER RETINA MODEL

As described in the Introduction the retina possesses a multi-layer structure where each layer contains different cell types with specific functions. This distinctive structure can be captured best through a model which exhibits a similar deep architecture and utilises multiple abstraction levels. Deep belief networks are highly suitable for extracting successive layers of representation, where features on each consecutive layer are of increasing complexity. We, therefore, promote the use of multi-layer deep networks for retina modelling purposes owing to their ability to extract a hierarchy of distinctive features from data and provide the required flexibility for modelling in an uncertain environment.

Our experiments investigate a deep belief network based model of early visual processing incorporating both the outer and inner plexiform layers of the retina. The weights of the network are learnt using the training algorithm of Hinton et al. (2006).

Simulated Environment

The retinal network implements the first stages of object recognition and visual information processing in general, by performing a variety of image enhancement and feature extraction routines. As discussed before, the majority of ganglion cells with their centre-surround receptive fields implement edge detection mechanisms similar to DoG filters. Here we show that deep networks, even when trained fully unsupervised on a simulated photoreceptor input succeed in learning DoG type feature detectors exhibiting centre-surround receptive fields. Examples of learnt features are shown in figure 4.

We model the input that reaches a tiny area of the retina as circular spots of different size and colour on a uniform background. To approximate different wavelength light input, our dataset contains images with various background colours showing circular spots which can overlap each other. We use RGB images in order to keep similarity with the photoreceptor input of trichromats. The images can be categorised into positive and negative classes

based on whether they contain a circle centred at the middle of the image. Examples are shown in the first rows of figure 3a and 3b. Our network model is first trained unsupervised to extract key features from this data. These features are subsequently utilised during the supervised training phase for the classification of images into two classes. This blob detection task mimics how ganglion cells learn to signal when differing light patterns reach the centre and surround of their receptive fields. We show in order to solve this task the network develops feature detectors similar to those implemented in the earliest stages of visual processing, including analogues of retinal ganglion cells with centre-surround receptive fields.

Figure 1b provides a schematic diagram of our DBN model trained on simulated photoreceptor input. As opposed to camera generated natural images or electrophysiological data from experiments, this simulated input and the corresponding class labels can be obtained with no cost and provides the advantage of having good control over the quality of data.

Training Deep Belief Networks

The key steps of the deep belief network training algorithm correspond to the unsupervised pretraining phase whereby the multi-layer representation is learnt one layer at a time using an RBM on each layer, followed by fine-tuning using backpropagation or a variant of the “wake-sleep” algorithm (Hinton et al., 2006). The latter is illustrated in figure 1b. The restricted Boltzmann machine probabilistic graphical model is based on the Boltzmann machine model which contains 2-layers of nodes: visible and hidden nodes and has no constraints on which nodes can be connected. RBMs on the other hand do not contain links between nodes on the same layer, hence the “restriction”. They are therefore bi-partite graphs and are significantly quicker to train due to the conditional independence of hidden nodes given the visible nodes, and vice versa. In our visual recognition task visible nodes correspond to image coordinates and hidden nodes represent image features. See figure 1b for illustration.

The probability of a configuration (state) of the visible and hidden nodes (v, h) can be calculated from the energy function of the RBM, which takes the form:

$$E(v, h) = -a'v - b'h - h'Wv, \quad (1)$$

where W is the weight matrix describing the connections between visible and hidden nodes, while a and b are the biases of the visible and hidden nodes respectively. The probability of a configuration is then given by:

$$p(v, h) = \frac{e^{-E(v, h)}}{\sum_{\eta, \mu} e^{-E(\eta, \mu)}}. \quad (2)$$

During the training phase, the probability of a given training example can be increased (the energy reduced) by altering the weights and biases. The following learning rule can be applied to maximise the log probability of

the training data:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}), \quad (3)$$

where ϵ is the learning rate and $\langle \cdot \rangle_{\phi}$ is used to denote expectations under the distribution ϕ .

Due to the conditional independence properties, sampling from $\langle v_i h_j \rangle_{data}$ is easy. In the case of RBMs with binary nodes, the probability of a hidden node h_j being 1 given a randomly chosen training image v is:

$$p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{ij}), \quad (4)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic sigmoid function. An example of an unbiased sample is then given by $v_i h_j$. Sampling for the visible nodes is similarly easy: the probability of a visible node v_i being 1 given the states of the hidden vectors is:

$$p(v_i = 1|h) = \sigma(a_i + \sum_j h_j w_{ij}). \quad (5)$$

On the other hand, sampling from $\langle v_i h_j \rangle_{model}$ is difficult and therefore approximations are normally applied. An efficient training method for RBMs which only broadly approximates the gradient of the log probability of the training data, is the single-step version (CD_1) of the Contrastive Divergence (Hinton, 2002) algorithm (CD). Each step of the algorithm corresponds to one step of alternating Gibbs sampling, where the states of the visible nodes are first set to a training example. Based on the states of the visible variables, binary states for the hidden nodes are sampled according to equation 4. This configuration of the hidden variables is subsequently used in the reconstruction phase where states for the visible nodes are sampled according to equation 5. The learning rule for the weights when using CD is then given by:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{reconst}) \quad (6)$$

and a similar rule can be applied for learning the biases.

In order to obtain an improved model, the sampling stage in each step can be continued for more iterations resulting in the general form of the CD algorithm: CD_n , where n is the number of alternating Gibbs sampling iterations. RBMs and stacked RBMs (DBNs) trained efficiently using CD are very powerful tools for learning generative models of visual or other types of complex data. If the data is continuous valued, such as in our case, visible nodes other than binary can provide better models. With this in mind we construct our networks using Gaussian visible nodes which are suitable for modelling our data. In the case of Gaussian visible nodes the energy function becomes:

$$E(v, h) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}, \quad (7)$$

where σ_i is the standard deviation corresponding to visible node i . For more details on training DBNs see (Hinton, 2012).

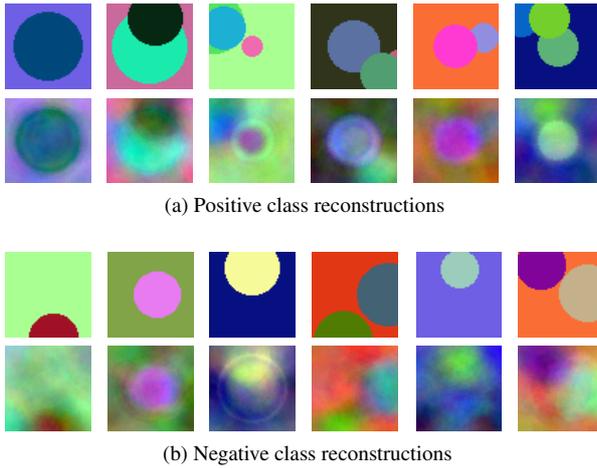


Figure 3: Figures illustrating how randomly selected examples of the test set are reconstructed by a deep network with 5 hidden layers. (a) and (b) show the reconstruction of positive (images with circle centred in the middle) and negative examples (images without a circle centred in the middle) respectively. The first row in both (a) and (b) corresponds to the test examples while the second row shows their reconstructions.

EXPERIMENTS

As outlined in the previous section, we model the earliest stages of visual processing using DBNs and test our model on a circular spot detection task using simulated photoreceptor input.

Dataset

Our dataset is divided into a training and a test set, containing 10 000 and 5 000 64x64x3 RGB images respectively. Ten different background colours can be found in the training set and five in the test set. The background colours of the training and the test set are different in order to ensure substantial dissimilarity between training and test samples. Examples are shown in the first rows of figure 3a and 3b.

Training

All experiments were performed with a variety of different settings, including different numbers of hidden nodes per layer and different learning rate values. From these settings only the top performing ones are discussed here. On each layer an RBM with Gaussian visible nodes and binary hidden nodes was used for pretraining. The pretraining consisted of 10 000 epochs and in each epoch all examples of the training data was used for updating the RBM parameters. Subsequently, for the classification task 200 epochs of backpropagation was used for fine-tuning. Hidden node numbers of 100, 500 and 2 000 was examined and the network depth was ranging between 1 to 5 hidden layers.

Testing

We investigate the effectiveness of the unsupervised training phase by visualising the features learnt by RBMs

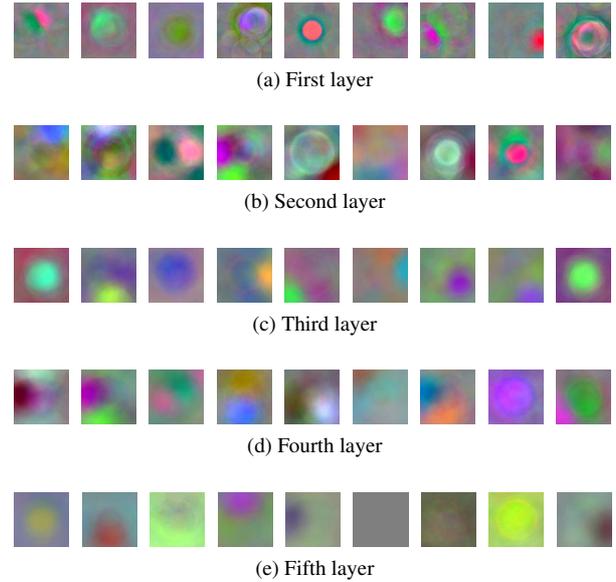


Figure 4: Random samples of features learnt on consecutive layers during the unsupervised pretraining phase in a deep network with 5 hidden layers (500 nodes per layer). Features on higher layers are visualised by linear combination of previous features. The learnt features contain a number of DoG detectors and Gaussian derivatives.

on each layer of the network. As the weights of the first layer hidden nodes correspond to image locations, the visualisation of first layer features can be obtained by displaying the weight vector of the hidden node in the shape of the original image. Visualising the consecutive layers is not straightforward due to the non-linearities and only approximate features can be shown. We chose to show the common visualisation method of higher level features obtained by linear combination of features from the previous layers which makes it easy to see receptive field areas.

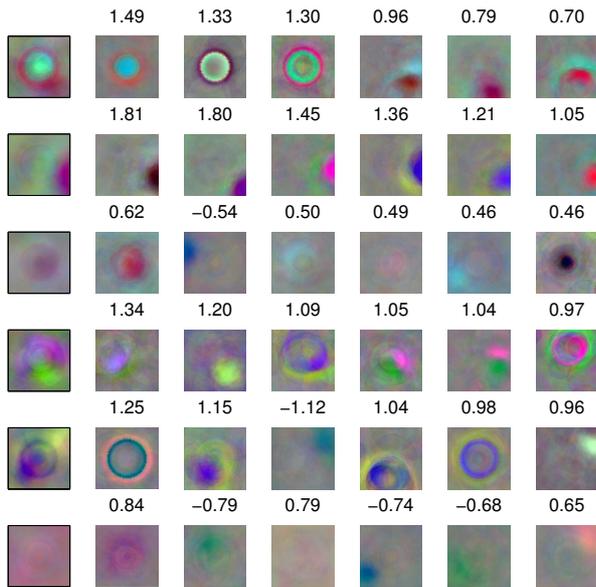
We also examine reconstructions of test examples calculated by the generative model. Reconstructions are obtained by feeding in an example to the network and subsequently calculating the top-down activations.

After the pretraining phase the network is trained by backpropagation to classify the input into 2 classes based on the existence of a circle in the centre of the image. Testing of the classifier is conducted on the test set. The change in performance of the classifier during the epochs of the backpropagation is measured by calculating the precision P , recall R and F-measure scores $F = 2PR/(P + R)$.

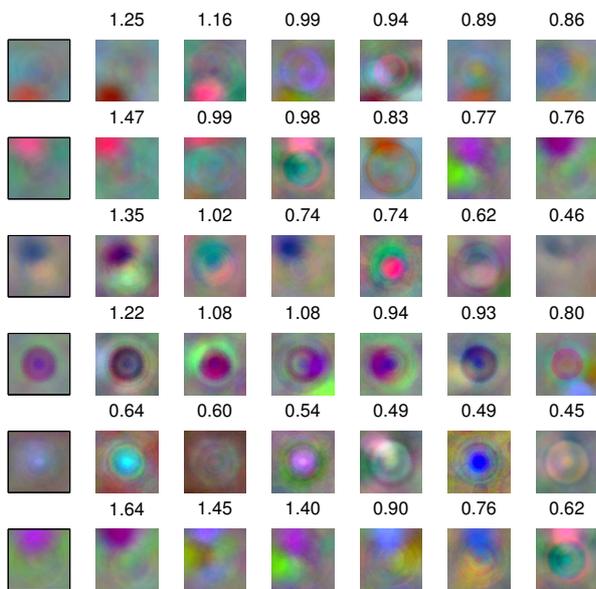
EVALUATION

Reconstruction

Positive and negative class samples of the test set are shown in figure 3 together with their reconstructions generated by a 5 hidden layer deep network model after the unsupervised training phase. As can be seen the DBN generative model learns to encode the data with a limited

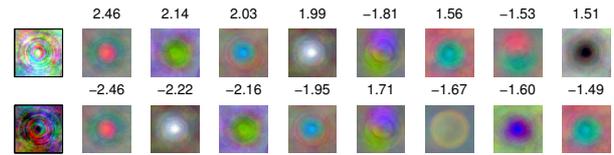


(a) Second - First layer



(b) Third - Second layer

Figure 5: Figure showing how features from a lower layer are combined to form features on the subsequent layer. In (a) and (b) the first image in each row shows the feature detector learnt by a randomly selected node on the higher layer. The consecutive images in the row represent features from the lower layer which have the highest corresponding weights to the higher layer feature. These features are sorted according to the strength of their connections, decreasing from left to right. The corresponding weight is shown above each image. The visualisation of high level features is given by a linear combination of lower level features with their corresponding weights. (a) shows how second layer features are built up of first layer features, while (b) illustrates how third layer features are composed using second layer features. Note the RBMs often group similar lower level features to form a higher level feature.



(a) Output - Third layer

Figure 6: Visualisation of the output layer and its connections to the highest level features in a network with 3 hidden layers trained using layer-by-layer pretraining followed by backpropagation. The first image in each row corresponds to an output layer node. The consecutive images in the row, similarly to figure 5, show features from the previous layer that have the strongest connections to the given output node. The corresponding weight is shown above each image. The first row visualises the strongest features and their weights for the positive class label (images with circle centred in the middle), while the second row corresponds to the negative class label.

number of nodes in a way that key features of even the unseen examples can be retained when reconstructed.

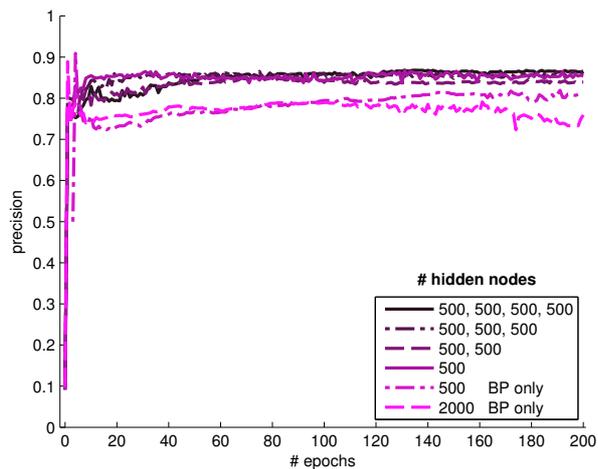
Classification

The precision, recall and the F-measure values achieved on the test set is shown in figure 7, for different numbers of hidden layers. These measures are plotted against the number of backpropagation epochs. One can see that the two networks trained using backpropagation without pretraining perform much worse than the pretrained networks. We noticed the features learnt by networks without pretraining were indistinct and noisy. The results also show multi-layered (2-4 hidden layers) networks perform superior to shallow networks, with 4 hidden layers being the best.

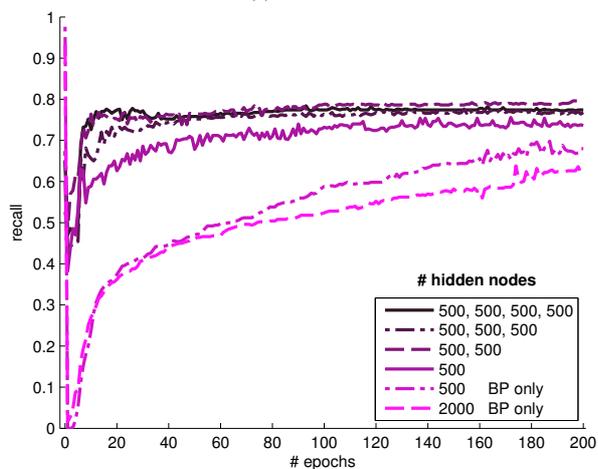
When trained in an adequate manner, each new layer of a DBN improves the generative model (Hinton et al., 2006). The strength of the multi-layered generative models is revealed when examining the graphs in figure 7: deep networks with at least 2 hidden layers can achieve high accuracy after only a few backpropagation epochs. This shows the unsupervised pretraining phase initialises the weights of the network to a favourable range and therefore less epochs of backpropagation is enough to guarantee good classification performance.

Features

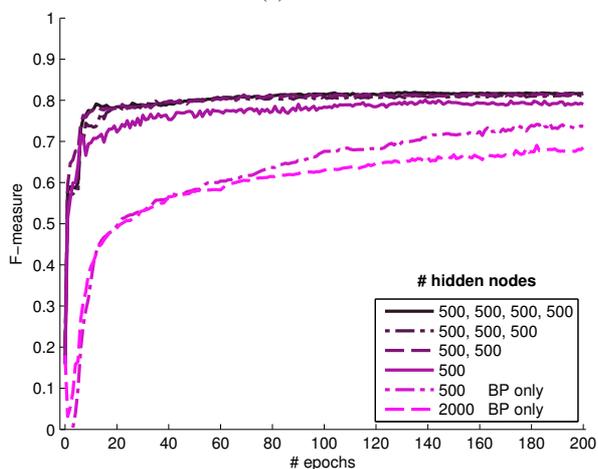
Figure 4 shows randomly sampled features learnt by RBMs on each layer of a 5 hidden layer network. Figure 5 and 6 show features after backpropagation in a 3 hidden layer network. For each randomly chosen higher level feature, those features from the previous layer are displayed which have the strongest connections to the higher level feature. The feature visualisation reveals the network succeeded in learning DoG detectors and a number of features typical to early visual processing in the retina and visual cortex. Similarity to ON/OFF ganglion cells with centre-surround receptive fields are clearly noticeable in many of the learnt feature detectors.



(a) Precision



(b) Recall



(c) F-measure

Figure 7: (a)-(c) show the change in precision, recall and the F-measure respectively on the test set (containing 5 000 samples) during the epochs of the backpropagation algorithm, when the number of hidden layers increases from 1 to 4. Each hidden layer has 500 hidden nodes, except for the last network containing 2 000 hidden nodes on 1 layer. In the first four cases the networks have been pretrained using RBMs before starting the backpropagation, as opposed to the last two cases (500 and 2 000 BP only) where only backpropagation was used for training.

This result could be anticipated as our input data and classification task of detecting spots in front of different coloured background, mimics functionality implemented in the early visual system. However, it is not obvious a DoG filter should emerge, as circles could be detected using different types of features. The fact that the network has learnt DoG filters in order to solve this task is therefore an important discovery.

Although the DBNs have been shown to have the ability to implement similar functionality to retinal ganglion cells, the underlying mechanisms, such as the network structure, is likely to be different. This is due to the fact that specific connections between retinal cells are not hard-coded into the algorithm, but are learnt in a primarily unsupervised fashion.

CONCLUSIONS

We have proposed a flexible probabilistic model of the retina and early stages of visual processing based on the state-of-the-art deep belief networks. We highlighted the resemblance between the inherently multi-layered retinal network and the deep network model. We trained our model on simulated photoreceptor input and evaluated the performance on a spot detection task, resembling functionality implemented in the early visual system. Supervised training of the networks with backpropagation was preceded by an unsupervised pretraining routine using RBMs. The multi-layer models achieved good classification results and, among other features of early visual processing, the networks learnt DoG feature detectors with centre-surround receptive fields resembling retinal ganglion cells.

In future work we will experiment with the deep network retina model using different types of input, e.g. natural image data, and different tasks corresponding to retinal ganglion cell and early visual processing functionalities. Furthermore, we will incorporate our retinal feature detectors into a convolutional network framework.

REFERENCES

- Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the International Conference on Machine Learning*.
- Eslami, S. M. A., Heess, N., and Winn, J. (2012). The shape Boltzmann machine: a strong model of object shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Gaudio, P. (1992). A unified neural network model of spatiotemporal processing in X and Y retinal ganglion cells. *Biological Cybernetics*, 67(1):23–34.
- Gollisch, T. and Meister, M. (2010). Eye smarter than scientists believed : Neural computations in circuits of the retina. *Neuron*, 65(2):150–164.
- Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

- Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 599–619. Springer Berlin Heidelberg.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.
- Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.
- Kameneva, T., Meffin, H., and Burkitt, A. N. (2011). Modelling intrinsic electrophysiological properties of ON and OFF retinal ganglion cells. *Journal of Computational Neuroscience*, 31(3):547–561.
- Kavukcuoglu, K., Sermanet, P., Boureau, Y. L., Gregor, K., Mathieu, M., and LeCun, Y. (2010). Learning convolutional feature hierarchies for visual recognition. In *Advances in Neural Information Processing Systems*.
- Kien, T. T., Maul, T., Ren, L. J., and Bargiela, A. (2012). Outer plexiform layer receptive fields as underlying factors of the Hermann grid illusion. In *Proceedings of the IEEE-EMBS International Conference on Biomedical Engineering and Sciences*.
- Kolb, H. (2003). How the retina works. *American Scientist*, 91(1):28–35.
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the International Conference on Machine Learning*.
- Le, Q. V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M. A., Dean, J., and Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *Proceedings of the International Conference on Machine Learning*.
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009a). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the International Conference on Machine Learning*.
- Lee, H., Largman, Y., Pham, P., and Ng, A. Y. (2009b). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*.
- Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, 76(2):266–280.
- Maul, T. H., Bargiela, A., and Ren, L. J. (2011). Cybernetics of vision systems: Toward an understanding of putative functions of the outer retina. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 41(3):398–409.
- Ölveczky, B. P., Baccus, S. A., and Meister, M. (2003). Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408.
- Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.
- Salakhutdinov, R. R. and Hinton, G. E. (2007). Using deep belief nets to learn covariance kernels for Gaussian processes. In *Advances in Neural Information Processing Systems*.
- Salakhutdinov, R. R. and Hinton, G. E. (2009). Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.
- Shlens, J., Rieke, F., and Chichilnisky, E. J. (2008). Synchronized firing in the retina. *Current Opinion in Neurobiology*, 18(4):396–402.
- Teeters, J., Jacobs, A., and Werblin, F. (1997). How neural interactions form neural responses in the salamander retina. *Journal of Computational Neuroscience*, 4(1):5–27.
- Usui, S., Ishihaiza, A., Kamiyama, Y., and Ishii, H. (1996). Ionic current model of bipolar cells in the lower vertebrate retina. *Vision Research*, 36(24):4069–4076.
- Velte, T. J. and Miller, R. F. (1997). Spiking and nonspiking models of starburst amacrine cells in the rabbit retina. *Visual Neuroscience*, 14(6):1073–1088.
- Wässle, H. (2004). Parallel processing in the mammalian retina. *Nature Reviews Neuroscience*, 5(10):747–757.

AUTHOR BIOGRAPHIES

DIANA TURCSANY is a PhD student in the School of Computer Science at The University of Nottingham since 2012. She obtained her MSc degrees from Eötvös Loránd University, Hungary in 2011 and from Cranfield University, UK in 2010. Her research interests lie within the areas of machine learning, neural computation and computer vision. Her current research focuses on developing computational models of the retina. Her email is dxt@cs.nott.ac.uk and her webpage is at <http://www.cs.nott.ac.uk/~dxt>.

ANDRZEJ BARGIELA is a Professor at The University of Nottingham. Since 1978 he has pursued research on processing of uncertainty in the context of modelling and simulation of physical and engineering systems. His current research centres around computational intelligence and involves mathematical modelling, information abstraction, parallel computing, artificial intelligence, fuzzy sets and neurocomputing. His email is abb@cs.nott.ac.uk and his webpage is at <http://www.bargiela.com/>.

TOMAS MAUL is an Associate Professor at The University of Nottingham Malaysia Campus. Prior to that he worked for two years at MIMOS Bhd. as a Senior Researcher in the fields of pattern recognition and computer vision. He obtained his PhD from the University of Malaya in Computational Neuroscience. Currently, he conducts research in the areas of neural computation, optimisation and computer vision. His email is Tomas.Maul@nottingham.edu.my and his webpage can be found at <http://kcztm.jupiter.nottingham.edu.my>.

PERFORMANCE COMPARISON OF EVOLUTIONARY TECHNIQUES ENHANCED BY LOZI CHAOTIC MAP IN THE TASK OF REACTOR GEOMETRY OPTIMIZATION

¹Michal Pluhacek, ¹Roman Senkerik, ¹Ivan Zelinka and ²Donald Davendra

¹Tomas Bata University in Zlin , Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{pluhacek,senkerik,zelinka}@fai.utb.cz

²Department of Computer Science, Faculty of Electrical Engineering and Computer Science
VB-TUO, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
donald.davendra@vsb.cz

KEYWORDS

PSO, Optimization, Evolutionary Algorithms, Swarm intelligence, Chaos

ABSTRACT

In this study the performance of two popular evolutionary computational techniques (particle swarm optimization and differential evolution) is compared in the task of batch reactor geometry optimization. Both algorithms are enhanced with chaotic pseudo-random number generator (CPRNG) based on Lozi chaotic map. It is depicted the course of the reactor processes dynamical parameters for the best obtained solutions for both algorithms and numerical values of cost functions are compared. The promising results are discussed.

INTRODUCTION

In recent years the demand for effective optimization tools for complex optimization tasks have significantly increased. For this reason a significant effort is put into development and enhancement of such methods. A group of very effective methods that had already established themselves are the evolutionary computational techniques (ECTs). Two of the most significant and potent (meta)heuristic algorithms from the ECTs field are the Differential evolution (DE) (Storn, Price, 1997) and Particle swarm optimization (PSO) (Kennedy, Eberhart 1995, Eberhart, Kennedy 2001). In the recent research (Pluhacek et al. 2013, Senkerik et al. 2013), it has been observed that through utilization of chaos based pseudo-random number generators within these heuristics their performance can be significantly improved. Thus the chaos enhanced ECTs can be effectively used to solve complex practical tasks (Pluhacek et al. 2012). In this paper the performance of both aforementioned algorithms is investigated in the task of batch reactor geometry optimization. This task represents the complex optimization problems with very slow inner dynamic. Both algorithms are enhanced with Lozi chaotic map (Spratt 2003). The newly acquired results of chaos driven PSO are compared to previously published

results of chaos enhanced DE algorithm (Senkerik et al. 2013).

The rest of the paper is structured as follows: In the next section the PSO algorithm is described in details followed by a brief description of the Lozi chaotic map that was used as the pseudo-random number generator (CPRNG) for both algorithms. In the next three sections the batch process, the batch reactor and the experiment setup are explained. The graphical and numerical results presentation and comparison follow afterwards. Finally the obtained results are briefly discussed in the last section that is followed by the conclusion.

PARTICLE SWARM OPTIMIZATION

The Particle Swarm Optimization algorithm (PSO) is the evolutionary optimization algorithm based on the natural behavior of bird and fish swarms and was firstly introduced by R. Eberhart and J. Kennedy in 1995 (Kennedy, Eberhart 1995, Eberhart, Kennedy 2001). As an alternative to genetic algorithms (Goldberg, David, 1989) and differential evolution (Storn, Price, 1997), PSO proved itself to be able to find better solutions for many optimization problems.

In the PSO algorithm the particles move through the multidimensional space of possible solutions. The new position of the particle in the next iteration is then obtained as a sum of actual position and velocity. The velocity calculation follows two natural tendencies of the particle: To move to the best solution found so far by the particular particle (known in the literature as personal best: *pBest* or local best: *lBest*). And to move to the overall best solution found in the swarm or defined sub-swarm (known as global best: *gBest*).

In the original PSO the new position of particle is altered by the velocity given by (1):

$$v_{ij}^{t+1} = w \cdot v_{ij}^t + c_1 \cdot Rand \cdot (pBest_{ij} - x_{ij}^t) + c_2 \cdot Rand \cdot (gBest_j - x_{ij}^t) \quad (1)$$

Where:

v_i^{t+1} - New velocity of the i th particle in iteration $t+1$.

w – Inertia weight value.

v_i^t - Current velocity of the i th particle in iteration t .

c_1, c_2 - Priority factors (set to the typical value = 2).

$pBest_i$ – Local (personal) best solution found by the i th particle.

$gBest$ - Best solution found in a population.

x_{ij}^t - Current position of the i th particle (component j of the dimension D) in iteration t .

$Rand$ – Pseudo random number, interval (0, 1). The chaotic pseudo-random number generator is applied here.

The maximum velocity of particles in the PSO is typically limited to 0.2 times the range of the optimization problem and this pattern was followed in this study. The new position of a particle is then given by (2), where x_i^{t+1} is the new particle position:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (2)$$

Finally the linear decreasing inertia weight (Nickabadi et al., 2011). is used in the PSO here. Its purpose is to slow the particles over time thus to improve the local search capability in the later phase of the optimization. The inertia weight has two control parameters w_{start} and w_{end} . A new w for each iteration is given by (3), where t stands for current iteration number and n stands for the total number of iterations.

$$w = w_{start} - \frac{((w_{start} - w_{end}) \cdot t)}{n} \quad (3)$$

CHAOTIC LOZI MAP

The Lozi map is a simple discrete two-dimensional chaotic map. The Lozi map is depicted in Fig. 1. The map equations are given in Eq. 4. Typical parameters given in literature (Sprott 2003) are $a = 1.7$ and $b = 0.5$.

$$\begin{aligned} X_{n+1} &= 1 - a|X_n| + bY_n \\ Y_{n+1} &= X_n \end{aligned} \quad (4)$$

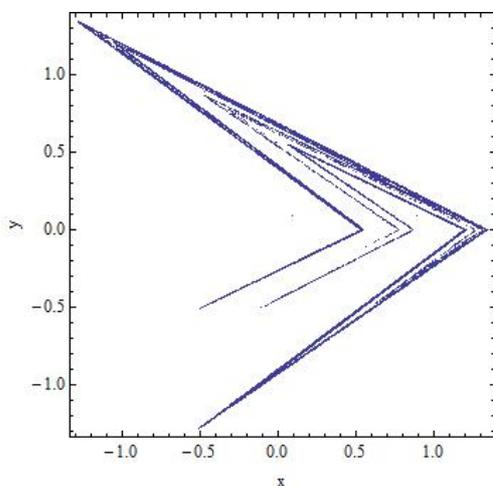


Figure 1: x, y plot of the Lozi map

BATCH PROCESSES

The optimization of batch processes has attracted attention in recent years (Silva and Biscaia 2003, Arpornwichanop et al 2005.). Batch and semi-batch processes are of considerable importance in the fine chemicals industry. A wide variety of special chemicals, pharmaceutical products, and certain types of polymers are manufactured in batch operations. In batch operations, all the reactants are charged in a tank initially and processed according to a pre-determined course of action during which no material is added or removed. From a process systems point of view, the key feature that differentiates continuous processes from batch and semi-batch processes is that continuous processes have a steady state, whereas batch and semi-batch processes do not (Srinivasan et al. 2002a, 2002b). Many modern control methods for chemical reactors were developed embracing the approaches such as iterative learning model predictive control (Wang et al. 2008), iterative learning dual-mode control (Cho et al. 2009) or adaptive exact linearization by means of sigma-point kalman filter (Beyer et al. 2008). Also the fuzzy approach is relatively often used (Sarma 2001). Finally the methods of artificial intelligence are very frequently discussed and used. Many papers present successful utilization of either neural networks (Sjöberg and Mukul 2002, Mukherjee and Zhang 2008, Mujtaba et al. 2006 or genetic (evolutionary) algorithms (Causa et al. 2008, Altinten et al. 2008, Faber et al. 2005).

This paper presents the static optimization of the batch reactor geometry by means of chaos driven PSO algorithm with CPRNG based on Lozi chaotic map and further compares the results with previously published results of similar optimization by means of chaos enhanced Differential evolution (Senkerik et al. 2013).

DESCRIPTION OF THE REACTOR

This research uses the mathematical model of the reactor shown in Fig. 2. The reactor has two physical inputs: one for chemical substances Chemical FK (Filter Cake) with parameters temperature- T_{FK} , mass flow rate- \dot{m}_{FK} and specific heat- c_{FK} ; and one for cooling medium of temperature T_{VP} , mass flow rate- \dot{m}_v and specific heat- c_v . The reactor has one output: cooling medium flows through the jacket inner space of the reactor, with volume related to mass - m_{VR} , and flows out through the second output, with parameters mass flow rate m_v , temperature- T_v and specific heat- c_v .

At the beginning of the process there is an initial batch inside the reactor with parameter mass- m_p . The chemical FK is then added to this initial batch, so the reaction mixture inside the reactor has total mass- m , temperature- T and specific heat- c_R , and also contains partially unreacted portions of chemical FK described by parameter concentration a_{FK} . In general, the reaction is highly exothermic, thus the most important parameter is the temperature of the reaction mixture. This temperature must not exceed 100°C because of safety aspects and quality of the product. The original design

of the reactor was based on standard chemical-technological methods and gives a proposal of reactor physical dimensions and parameters of chemical substances. These values are called within this paper original parameters. For the detailed description, please refer to (Senkerik and Zelinka 2005).

The main objective of the optimization is to achieve the processing of large amount of chemical FK in a very short time.

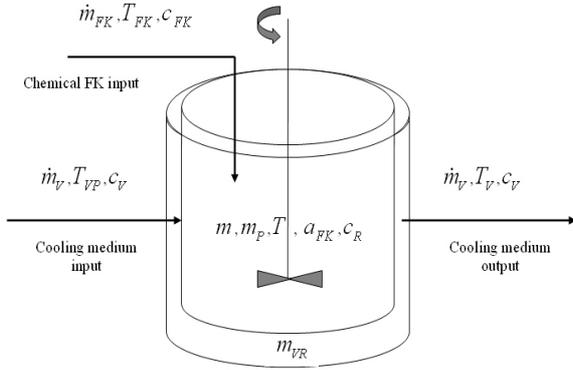


Figure 2: Batch reactor

Description of the reactor applies a system of four balance equations (5) and one equation (6) representing the term “ k ”.

$$\begin{aligned} \dot{m}_{FK} &= m'[t] \\ \dot{m}_{FK} &= m[t] a'_{FK}[t] + k m[t] a_{FK}[t] \\ \dot{m}_{FK} c_{FK} T_{FK} + \Delta H_r k m[t] a_{FK}[t] &= \\ &= K S (T[t] - T_V[t]) + m[t] c_R T'[t] \\ \dot{m}_V c_V T_{VP} + K S (T[t] - T_V[t]) &= \dot{m}_V c_V T_V[t] + m_{VR} c_V T'_V[t] \end{aligned} \quad (5)$$

$$k = A e^{-\frac{E}{RT[t]}} \quad (6)$$

EXPERIMENT SETUP

The control parameters of PSO algorithm were set following way:

Population size: 50
 Iterations: 200
 w_{start} : 0.9
 w_{end} : 0.4
 v_{max} : 0.2

The control parameters of DE algorithm were set as follows:

Population size: 50
 Generations: 200
 F : 0.8
 CR : 0.8

RESULTS COMPARISON

In this section the results of chaotic PSO algorithm are presented and compared with the results of chaotic Differential evolution (ChaosDE). Both algorithms used CPRNG based on Lozi Chaotic map. Table 1 contains the overview of optimized and original parameters where the internal radius of reactor is expressed by parameter r and is related to cooling area S . Parameter d represents the distance between the outer and inner jackets and parameter h means the height of the reactor. The original design of the reactor was based on standard chemical-technological methods and gives a proposal of reactor physical dimensions and parameters of chemical substances. These values are called within this paper *original parameters (values)*. The courses of the reactor processes dynamical parameters for the best obtained solutions for both PSO and DE are compared in figures 3 – 8. Table 2 contains the simple statistical evaluation of the repeated runs of both algorithms.

Table 1: Optimized reactor parameters, difference between original and the optimized reactor

Parameter	Range	Original value	Optimized value (PSO)	Optimized value (DE)
\dot{m}_{FK} [kg.s ⁻¹]	0 – 10.0	0 - 3	0.06732	0,06935
T_{VP} [K]	273.15 – 323.15	293.15	273.664	276.819
\dot{m}_V [kg]	0 – 10.0	1	3.7451	4,9832
r [m]	0.5 – 2.5	0.78	1.3896	1.0008
h [m]	0.5 – 2.5	1.11	1.0668	1.0088
d [m]	0.01 – 0.2	0.03	0.091965	0.0216

Table 2: Statistical overview and comparison

Algorithm	Avg CF	Median CF	Max CF	Min CF	StdDev
Chaos PSO with Lozi Map	26792.1	28557.4	31615.9	21560.3	3991.73
Chaos DE with Lozi Map	21117.64	21113.13	21232.46	21046.02	40.4

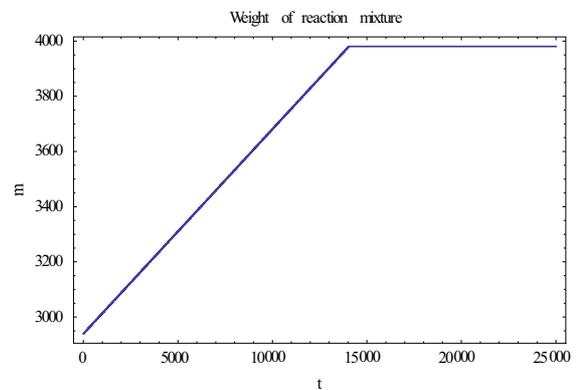


Figure 3: Weight of reaction mixture (PSO)

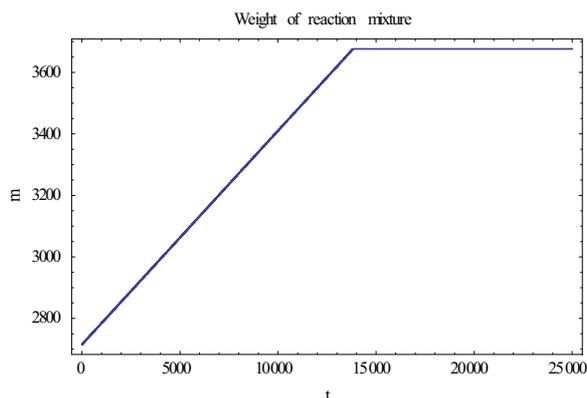


Figure 4: Weight of reaction mixture (DE)

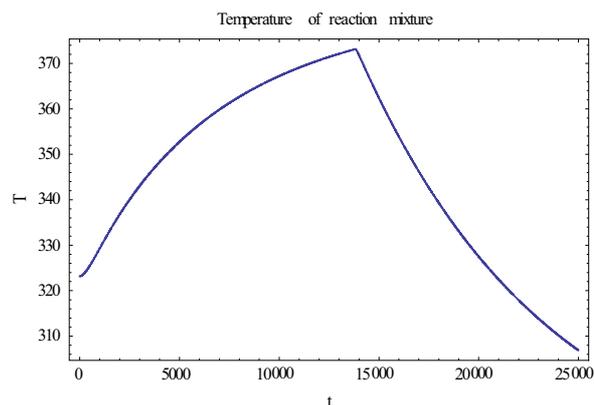


Figure 8: Temperature of reaction mixture (DE)

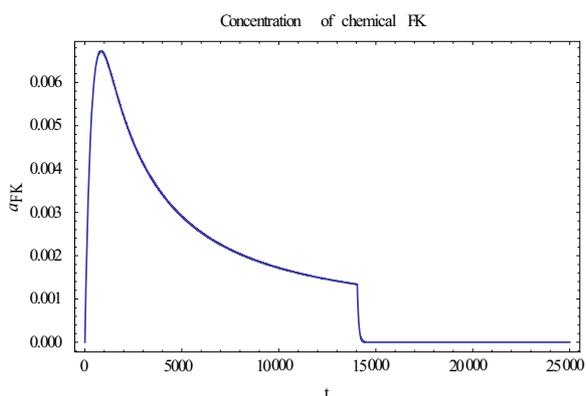


Figure 5: Concentration of chemical FK (PSO)

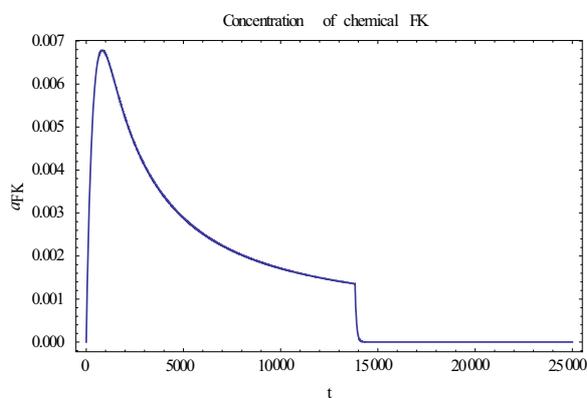


Figure 6: Concentration of chemical FK (DE)

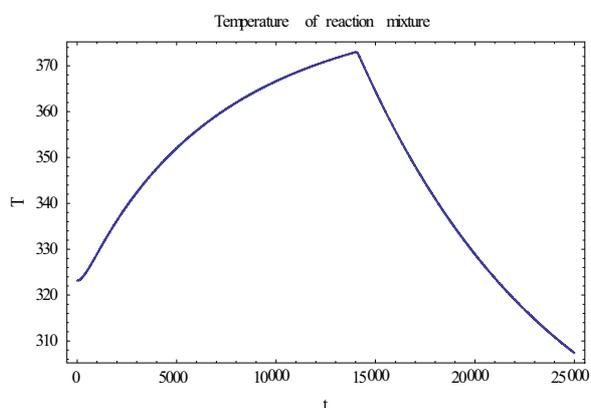


Figure 7: Temperature of reaction mixture (PSO)

BRIEF ANALYSIS OF THE RESULTS

Compared results in Table 2 indicate that the DE algorithm enhanced with Lozi chaotic maps seems to perform better on the task of reactor geometry optimization than chaotic PSO algorithm. However the evaluation of best found solutions (Table 1) depicted in figures 3 – 8 seems to indicate that the difference in final batch process behavior is not very perceptible. Both algorithms managed to find successful solution as none of the critical values were exceeded. The concentrations of chemical substances and weights of reaction mixture are comparable, as well as the courses of temperature.

CONCLUSION

In this paper the performance of PSO algorithm driven by Lozi chaotic map was compared to the performance of Differential evolution with similar chaos based PRNG in the task of batch reactor optimization. The final solution was evaluated and compared. Despite the differences in final cost function value in favor of the DE algorithm, it can be stated that both algorithms managed to obtain successful reactor designs. The chaos driven evolutionary computational techniques seem to be very promising tools for optimization of very complex and slow dynamic processes such as the task presented in this work and future research shall investigate this further.

ACKNOWLEDGEMENT

The following grants are acknowledged for the financial support provided for this research: Grant Agency of the Czech Republic - GACR P103/13/08195S, is partially supported by Grant of SGS No. SP2014/159 and SGS No. SP2014/170 VŠB - Technical University of Ostrava, Czech Republic, by the Development of human resources in research and development of latest soft computing methods and their application in practice project, reg. no. CZ.1.07/2.3.00/20.0072 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic, further

was supported by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2014/010.

REFERENCES

- Altintin A., Ketevanlioglu F., Erdogan S., Hapoglu H. and Albaz M., "Self-tuning PID control of jacketed batch polystyrene reactor using genetic algorithm", *Chemical Engineering Journal*, vol. 138, 2008, pp. 490-497.
- Arpornwichanop A., Kittisupakorn P. and Mujtaba M.I., "On-line dynamic optimization and control strategy for improving the performance of batch reactors". *Chemical Engineering and Processing*, vol. 44 (1), 2005, pp. 101-114.
- Aydin I., Karakose M. and Akin E., "Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection", *Expert Systems with Applications*, vol. 37, no. 7, 2010, pp. 5285-5294.
- Beyer M. A., Grote W., Reinig G., "Adaptive exact linearization control of batch polymerization reactors using a Sigma-Point Kalman Filter", *Journal of Process Control*, vol.18,2008, pp. 663-675.
- Causa J., Karer G., Nunez A., Saez D., Skrjanc I. and Zupancic B., "Hybrid fuzzy predictive control based on genetic algorithms for the temperature control of a batch reactor". *Computers and Chemical Engineering*, vol. 32 (8), 2008, pp. 3254 – 3263.
- Cho W., Edgar T.F. and Lee J., "Iterative learning dual-mode control of exothermic batch reactors", *Control Engineering Practice*, vol 16, 2008, pp. 1244-1249.
- Dorigo, M., *Ant Colony Optimization and Swarm Intelligence*, Springer, 2006.
- Eberhart, R., Kennedy, J., *Swarm Intelligence, The Morgan Kaufmann Series in Artificial Intelligence*, Morgan Kaufmann, 2001.
- Faber R., Jockenhövel T. and Tsatsaronis G., "Dynamic optimization with simulated annealing", *Computers and Chemical Engineering*, vol. 29, 2005, pp. 273-290.
- Goldberg, David E. (1989). *Genetic Algorithms in Search Optimization and Machine Learning*. Addison Wesley. p. 41. ISBN 0201157675.
- Kennedy, J.; Eberhart, R. (1995). "Particle Swarm Optimization". *Proceedings of IEEE International Conference on Neural Networks*. IV. pp. 1942-1948
- Liang W., Zhang L. and Wang M., "The chaos differential evolution optimization algorithm and its application to support vector regression machine", *Journal Of Software*, vol. 6, no. 7, 2011, pp. 1297-1304.
- Mujtaba M., Aziz N. and Hussain M.A., "Neural network based modelling and control in batch reactor", *Chemical Engineering Research and Design*, vol. 84 (8), 2006, pp. 635-644
- Mukherjee A. and Zhang J., "A reliable multi-objective control strategy for batch processes based on bootstrap aggregated neural network models", *Journal of Process Control*, vol. 18, 2008, pp. 720-734.
- Nickabadi A., Mohammad Mehdi Ebadzadeh, Reza Safabakhsh, A novel particle swarm optimization algorithm with adaptive inertia weight, *Applied Soft Computing*, Volume 11, Issue 4, June 2011, Pages 3658-3670, ISSN 1568-4946
- Pluhacek M., Senkerik R, Davendra D., Kominkova Oplatkova Z., and Zelinka I., "On the behavior and performance of chaos driven PSO algorithm with inertia weight," *Computers & Mathematics with Applications*, vol. 66, pp. 122-134, 2013.
- Pluhacek M., Senkerik R., Davendra D., Zelinka I. PID Controller Design For 4th Order system By Means Of Enhanced PSO algorithm With Lozi Chaotic Map, In: *Proceedings of 18th International Conference on Soft Computing - MENDEL 2012*, 2012, pp. 35 - 39, 2012, ISBN 978-80-214-4540-6.
- Sarma P., "Multivariable gain-scheduled fuzzy logic control of an exothermic reactor", *Engineering Applications of Artificial Intelligence* vol. 14, 2001, pp. 457-471.
- Senkerik R. and Zelinka I., "Optimization and control of batch reactor by evolutionary algorithms", In *Proceedings of 19th European Conference on Modelling and Simulation - ECMS*, 2005, pp. 59-65.
- Senkerik, R.; Pluhacek, M.; Oplatkova, Z.K.; Davendra, D.; Zelinka, I., "Investigation on the Differential Evolution driven by selected six chaotic systems in the task of reactor geometry optimization," *Evolutionary Computation (CEC)*, 2013 IEEE Congress on , vol., no., pp.3087,3094, 20-23 June 2013 doi: 10.1109/CEC.2013.6557946
- Shi Y.H., Eberhart R.C., A modified particle swarm optimizer, *IEEE International Conference on Evolutionary Computation*, Anchorage Alaska, 1998, pp. 69-73
- Silva C.M. and Biscaia E.C., "Genetic algorithm development for multi-objective optimization of batch free-radical polymerization reactors", *Computers and Chemical Engineering*, vol. 27, 2003, pp. 1329-1344.
- Sjöberg J. and Mukul A., "Trajectory tracking in batch processes using neural controllers", *Engineering Applications of Artificial Intelligence* vol. 15, 2002, pp. 41-51.
- Sprott J. C., "Chaos and Time-Series Analysis", Oxford University Press, 2003
- Srinivasan B., Palanki S. and Bonvin D., "Dynamic optimization of batch processes II: Role of Measurement in handling uncertainty". *Computers and Chemical Engineering*, vol. 27, 2002, pp. 27-44.
- Srinivasan B., Palanki S., and Bonvin D., "Dynamic optimization of batch processes I: Characterization of the nominal solution". *Computers and Chemical Engineering*, vol. 27, 2002, pp. 1-26.
- Storn R., Price K., Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization* 11 (1997) 341-359
- Wang Y., Zhou D. and Gao F., "Iterative learning model predictive control for multi-phase batch processes", *Journal of Process Control*, vol. 18, 2008, pp. 543-557.
- Zhenyu G., Bo Ch., Min Z., and Binggang C., "Self-Adaptive Chaos Differential Evolution", *Lecture Notes in Computer Science*, vol. 4221, 2006, pp. 972-975.

AUTHOR BIOGRAPHIES

MICHAL PLUHACEK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Information Technologies and obtained his MSc degree in 2011. He is now a Ph.D. student at the same university. His email address is: pluhacek@fai.utb.cz



ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2004, Ph.D. degree in Technical Cybernetics in 2008 and Assoc. prof. in 2013 (Informatics). He is now an Assoc. prof. at the same university (research and courses in: Evolutionary Computation, Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). His email address is: senkerik@fai.utb.cz



IVAN ZELINKA was born in the Czech Republic, and went to the Technical University of Brno, where he studied Technical Cybernetics and obtained his degree in 1995. He obtained Ph.D. degree in Technical Cybernetics in 2001 at Tomas Bata University in Zlin. Now he is a professor at TBU in Zlin and Technical University of Ostrava (research in: Artificial Intelligence, Theory of Information). Email address: ivan.zelinka@vsb.cz.



ANALYSIS OF EEG SIGNAL FOR USING IN BIOMETRICAL CLASSIFICATION

Roman Zak, Jaromir Svejda, Roman Senkerik and Roman Jasek
Department of Informatics and Artificial Intelligence
Tomas Bata University in Zlin
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic.
E-mail: {rzak, svejda, senkerik, jasek}@fai.utb.cz

KEYWORDS

Brain Computer Interface, EEG, Signal processing, Neural network classification,

ABSTRACT

Aim of this article is to clarify the potential use of EEG signal in modern information age. The basic principle of Brain Computer Interface (BCI) lies in the connection of brain waves with output device through some interface. BCI technology represents a communication interface between brain and computer. To sense electric signal from the brain, it is usually used an equipment based on the last results of scientific research on neuro-technology. Communication is provided by wireless transmission through which the signal is transmitted from the equipment to personal computer. Then the signal is analysed, processed and used for finding appropriate classification parameters.

INTRODUCTION

Many scientific disciplines deal with the human brain; for example numerical neuroscience, neuro-informatics, informatics or medicine. All of them bring theories, which could explain different brain activities. Numerical neuroscience provides mathematical and biophysical models, which are able to model basic processes in neurons and neural networks. The main goal of neuro-informatics is systematic development of database intended to collect information such as brain morphology, brain parts anatomy and their functional connection, brain electrophysiology, brain states obtained with magnetic resonance and their integration. Further, it seeks to develop tools for modelling, where the aim is the most accurate emulation of brain activity. In Informatics, complex networks are highly suitable to model a complex system among which the brain includes. The contribution of medicine is undisputable especially in brain anatomy research.

The human brain is a complex system, which is an object of our research. It is regarded as the most complex system in the universe. The modern science is currently attempting to understand the complex interconnection among individual parts of the brain (Sporns et.al. 2005). There are many publications, which deal with description of the brain (Adeli 2010; Damasio 1995; Sporns et al. 2005).

Currently there are many known applications of BCI technology, but not enough at each particular field of study. Signal that is sensed from the brain is the key element in the BCI model; therefore the design of an appropriate algorithm for processing of the signal is the most discussed part of BCI model structure (Schalk et al. 2004).

Invasive methods of sensing the brain activity could provide very accurate data, but it is not both technically and user friendly; thus, it would not be further mentioned in this article. On the other hand, more accessible non - invasive methods can obtain relatively weak signal with amplitude ranging from units to hundreds of microvolts. Moreover, the signal is also prone to noise elements. Another disadvantage of this method is a summation of neuron signals; thus, obtained data are referenced to a specific group of neurons. The brain itself is composed of several parts, without which his activity could not be possible. One of its basic structural parts is a neuron. The neuronal cells are characterized by the fact that electrical activity is carried out in them. These cells communicate with each other by electrical signals. According to the last estimate, there are approximately 10^{11} neurons in the brain. Every one of them is connected with thousands of other neurons. The main source of EEG signal is an electric activity of synapse - dendrites membrane located in the surface layer of the cortex. Each active synapse dispatches electromagnetic pulse to the environment during excitation. Due to the high number of these pulses, it is difficult to locate their source by means of multichannel sensor on the skin. This issue could be compared to full amphitheatre, in which there are chanting people and the task is to recognize from outside, which specific group of fans shouts. A different perspective on this issue may be such that the aim is to identify uniqueness of the signal for each individual subject. In the example shown above, it is as we would like to recognize the type of the stadium by the mass of chanting people. For example, there is noticeable difference between hockey and tennis fans. The biometric signatures are different for each creature on the planet Earth.

METHODS

There are several approaches for sensing brain activity. The most widely used is EEG technology, which

belongs among the non – invasive methods. Devices based on EEG technology provide signal with very low voltage amplitude, because the signal has to pass through the relatively low conductive skull. The amplitude ranges from tens to hundreds microvolts.

Sensing the brain activity

Recently, we have used Emotiv EPOC neuroheadset to obtain EEG signal from the human brain. Sensing of EEG by Emotiv EPOC neuroheadset has a number of advantages, because it already involves solved elementary issues in the processing of the measured signal. Due to this fact, it is not necessary to operate with raw data. It depends on the further usage of the data. Although the spectrum of this data could be used in many applications, it is not simple to understand the entire significance of the whole signal even if the proportion of the noise is minimal. This technology has the greatest expansion and certainly also the priority significance in diagnosis of various diseases in medicine (Adeli 2010).

Emotiv Corporation developed personal brain - computer interface for human – computer interaction using neuro-technology, which is based on processing of electromagnetic waves occurring in human brain. The interface has wide range of possible applications; for example in interactive games, intelligent adaptive environment, audio visual art and design, medicine, robotics and automotive industry. Moreover, it can be deployed in large amount of scientific research.

Emotiv EPOC neuroheadset (Figure 1) measures a signal wirelessly transferred to common personal computer. It is a device, which has a set of sensors intended for sensing the activity produced by human brain. Traditional EEG devices requires the use of conductive pasta to improve the conductivity between electrodes and hairs. On the other hand, the neuroheadset do not need any additional tools. It has 14 high resolution sensors, which are placed on optimal positions on the human head (Figure 2).



Figure 1: Emotiv EPOC neuroheadset (Emotiv 2012)

Moreover, it also includes gyroscope for determinate the position in the area. Each channel has its own label

based on its position on the head: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. Sampling frequency of the neuroheadset is 2048 Hz. More information about neuroheadset can be found in (Emotiv 2012).

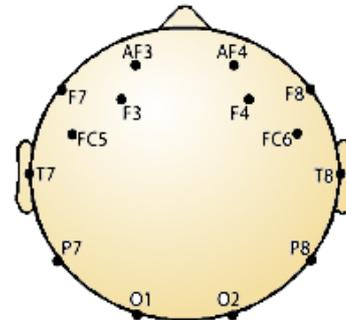


Figure 2: Placement of electrodes of Emotiv EPOC neuroheadset

Emotiv provide basic software set containing many tools, which can be used for recording various signals such as electric potential from all 14 sensors, power spectrum of individual EEG channels in real time and rotational acceleration of the head in horizontal and vertical axis using data from gyroscope. All of these outputs are shown in graphs. Data are also available in raw form, which can be used for further analysis. If it is required special functionality, which is not provided by native software, it is desirable to develop own application using Emotiv SDK (Software Development Kit).

Native software consists of three classification suites. Each of them enables the usage of algorithm developed by Emotiv. First of them is Expressive suite, which contains identification system for recognition of facial expression such as smile, eyewink, etc. The muscle signals are used for this purpose. The sources for these signals are obtained by sensors, which are located around the face.

The second suite can be used to measure and identification of emotional state; for example nervousness, alertness, concentration, etc. Therefore, it is called Affective suite. Muscle signals and ocular signals are filtered by specially designed filters; thus, identification algorithm uses clear brain signal.

The last suite is called Cognitive. This classification mode uses whole measured signal, which contains both clear brain signal and muscle signal. Classification algorithm is based on artificial intelligence methods. Type and structure of applied neural network is patented by Emotiv Corporation; therefore, the specific information about the algorithm is protected.

If it is required other processing of the signal than the native software allows, it may be processed by another software application.

Measured raw data can be subjected to offline analysis to research.

Processing the brain activity

If person could be recognized by custom EEG traces, it would mean that the person could be uniquely identified by EEG signal and it could bring new ways of authorization routines. Critical phase lies in signal classification. Even if the meaning of both the waveform and the signal content is not very important for classification purposes, there is another issue which have to be considered. EEG device provides large amount of data which has to be effectively and quickly processed in order to perform correct classification of the subject from the signal in real time. Classification tasks could be realized by using neural networks. However, it is difficult to predict, which neural network could use its cognitive potential for classification task mentioned above (Hazrati and Erfanian 2010). Investigation of the most appropriate classifier requires testing of many subjects. Furthermore, it is necessary to find algorithm with the shortest response time with respect to the credibility of obtained output. Another issue is to determinate which output is suitable. There could be considered the theory of large numbers; thus, maximum possible number of subjects needs to be tested. Therefore, it is more important to select key parameters of the signals that are different for each person. Even though the parameters are different for different people, the question remains whether the parameters remain constant in different time frames for the same person.

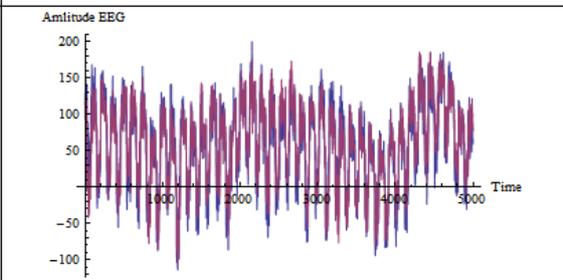
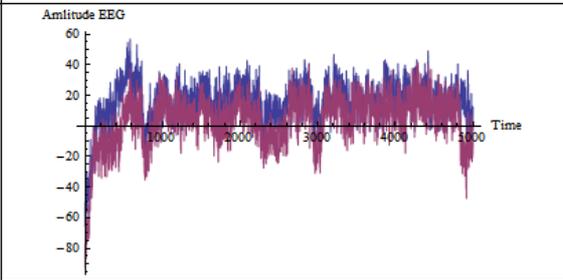
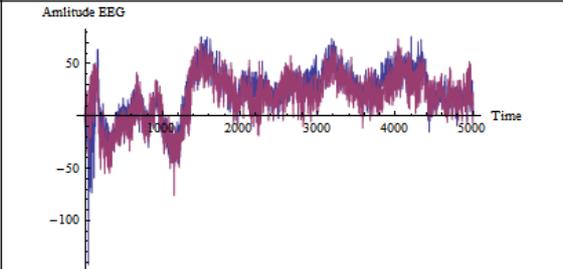
RESULTS

Idle state of mind was chosen as the optimal for the measurement. The relative idle state is such that the EEG signal does not contain any artefacts. Some artefacts are already filtered out before the signal processing. That mostly includes elements in signal that related to some physical responses such as eye blinking, motion and muscle activity, heartbeat etc. Furthermore, external artefacts interfering the signal are primarily eliminated in analog-to-digital conversion.

In order to perform classification, it has to be set unique characteristics of the signal (hereinafter called as classification parameters). That task is the first step of our research. The most important is to find appropriate classification parameters. However, the device returns fourteen channels with various amplitudes. Therefore, it is necessary to normalize the signal by applicable algorithm. This procedure count with sampling frequency of the neuroheadset, which is 128 Hz.

Prepared set of data is ready for another mathematical or statistical analysis. Firstly, it was performed a correlation analysis between channels of the first subject's EEG signal. All combinations of signals were tested in order to find which pair of channels influence each other. Further aim of the analysis was to find out whether another subjects have different relations between the channels. Correlation was calculated for each compared pair of channels.

Table 1: Example of data analysis

Person	Marginal match [%]	Name of channels	Correlation [%]	Correlated channels
Subject 1	31.8681	{AF3, F3}	95.7985	
Subject 2	2.1978	{O2, P8}	81.7761	
Subject 3	1.0989	{P7, O1}	87.9316	

Further, the same correlation threshold is set for all pairs. Then it is count the number of similarities moving over the threshold. Obtained value is converted into percentage (marginal match). Course of marginal match for 5000 samples is shown in Table 1. It can be observed that the marginal match is different for each subject. Even if the marginal match could be one of the appropriate classification parameters, it is necessary to confirm all complex biometric links on higher number of subjects. That remains as another object of our further research. However, good result is obvious from the waveform of the signal with the highest correlation. Further possible conclusion from data presented in Table 1 is that signals seem different even if they all had relatively similar initial conditions including closed eyes, no muscle motions and also the same time range. Further, it was investigated how the correlation changes with variety length of examined data. Figure 3 depicts three subjects compared in various offsets. The offset's length changes from 1000 to 30 000 with step 1000. It seems that a larger amount of data slightly decreases the correlation.

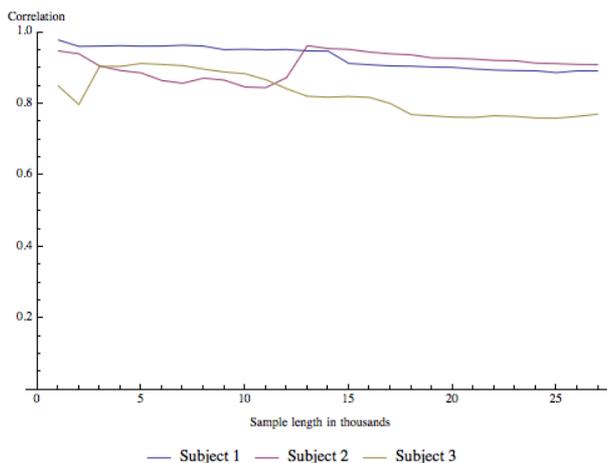


Figure 3: Dependence of correlation value to sample length

DISCUSSIONS

Human brain is the most complex known system in the universe. Study of its activity is extremely important mainly due to the more precise diagnosis of brain diseases and their treatment. Furthermore, acquired knowledge could be used in modern technologies with BCI systems, where an interaction between brain and computers appears.

We are currently performing the measurement of the EEG signal in our research. Our aim is to discover interesting regularities in the EEG signal waveform, which could contribute to the improvement of current approaches of brain activity simulation. Moreover, these regularities could be used to recognize some specific states of the brain, which can be then used to control the software or equipment connected to the computer.

There are many approaches to analysis of data. Moreover, EEG signal belongs to group of biometric signals which are usually very complex. The question remains whether it is possible to involve significance of all characteristics and signal history of EEG signal to classification process.

Biometrical data are typically represented as an image or a quantification of measured physiological or behavioural characteristics. As these data should refer to very complex human behaviour or describe very precisely physiological characteristic (typically iris scan, fingerprint, palm vein image, hand scan, voice, walk pattern etc.) these data can easily become very large and hard to process. For this reason a modern ways of data processing and classification are applied for biometrical data. The leading method is the usage of neural networks (Tangkraingkij 2009).

Correlation analysis demonstrates that there are relations between individual EEG channels. Further, it shows that the highest value of correlation was always found between neighbouring electrodes. Subject 1 has the highest correlation between electrode AF3 and F3 which are both located in frontal region of the brain. On the other hand, the subject 2 has the highest correlation between electrodes O2 and P8, which are located in rear regions of the brain. Both subject were measured in the idle state of mind. That behaviour of the subject's brains should be proven on more measurements of the same subjects. If the behaviour remained the same, it would mean, that it could be set as another classification parameter.

From obtained results we concluded that our future research could possibly answer the question which statistical characteristics are the most suitable for usage in classification algorithm based on neural network. For example, difference between individual subjects is the feature, which could be used as another classification parameter. Results described in this article are the first part of future extensive research.

ACKNOWLEDGEMENTS

This work was supported by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2014/31; further it was supported by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

REFERENCES

- Adeli H. 2010. *Wavelet-Chaos-Neural Network Models for EEG-Based Diagnosis of Neurological Disorders*. In: Kim T-h, Lee Y-h, Kang B-H, Slezak D (eds) *Future Generation Information Technology*, vol 6485. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 1-11. doi:10.1007/978-3-642-17569-5_1.
- Damasio, H. 1995. *Human brain anatomy in computerized images*. Oxford university press.

- Emotiv | EEG System | Electroencephalography. 2012. (online). Available from: <http://www.emotiv.com/index.php>.
- Hazrati, Mehrnaz Kh. a Abbas Erfanian. 2010. *An online EEG-based brain-computer interface for controlling hand grasp using an adaptive probabilistic neural network*. DOI: 10.1016/j.medengphy.2010.04.016.
- Schalk, G., D.J. Mcfarland, T. Hinterberger, N. Birbaumer a J.R. Wolpaw. 2004. *BCI2000: A General-Purpose Brain-Computer Interface (BCI) System*. DOI: 10.1109/TBME.2004.827072.
- Sporns, O., Tononi, G., Kötter, R. 2005. *The human connectome: a structural description of the human brain*. PLoS computational biology, 2005, 1.4: e42.
- Tangkraingkiij, P.; Lursinsap, C.; Sanguansintukul, S.; Desudchit, T. 2009. *Selecting Relevant EEG Signal Locations for Personal Identification Problem Using ICA and Neural Network*, Computer and Information Science. ICIS 2009. Eighth IEEE/ACIS International Conference on, On page(s): 616 – 621

AUTHOR BIOGRAPHIES

ROMAN ZAK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Information Technologies and obtained his MSc degree in 2011. He is now a Ph.D. student at the same university. His email address is: rzak@fai.utb.cz



JAROMIR SVEJDA was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Information Technologies and obtained his MSc degree in 2011. He is now a Ph.D. student at the same university. His email address is: svejda@fai.utb.cz



ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2004, Ph.D. degree in Technical Cybernetics in 2008 and Assoc. prof. in 2013 (Informatics). He is now an Assoc. prof. at the same university (Research and courses in: Applied Informatics, Cryptology, Artificial Intelligence, Mathematical Informatics). His email address is: senkerik@fai.utb.cz



USING ARTIFICIAL NEURAL NETWORK FOR THE KICK TECHNIQUES CLASSIFICATION – AN INITIAL STUDY

Dora Lapkova, Michal Pluhacek, Zuzana Kominkova Oplatkova, Milan Adamek

Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{dlapkova, pluhacek, oplatkova, adamek}@fai.utb.cz

KEYWORDS

Professional defense, Kick techniques, Direct kick, Round kick, Classification, Neural networks.

ABSTRACT

In this initial study it is investigated the possibility of using simple artificial neural network for classification of kick techniques based on their specific force course profile. The aim is to investigate whether the neural networks could be a suitable tool for such task and can be possibly used in following research that will deal with classification of punch techniques and also the striker's gender and level of training.

INTRODUCTION

The kick techniques are (apart from punching techniques) the most important and effective techniques in unarmed professional defense with significant force delivery. Various kick techniques are the subject of research investigation mostly for the needs of martial arts. (Liu et al. 2000, Pieter and Pieter 1995).

This paper presents initial results of analysis of two different kick techniques: the direct kick and the round kick (Liu et al. 2000). The aim was to find out whether it is possible to distinguish these two techniques from a kick impact force profile.

In this long-term research the participants were asked to perform a set of different punch and kick techniques on a measuring station. The impact force profiles were stored for further analysis. To uncover whether there are certain unique characteristics for the two kick techniques mentioned above the artificial neural network (ANN) was chosen as a suitable classifier.

Firstly, kick techniques are explained. In the following paragraph, measuring devices, the method of data storage and experiment setup for measurement are described. Artificial neural network theory is depicted in the next section. Problem definition and consequent analysis are followed by result section. The conclusion summarizes the kick techniques classification.

KICK TECHNIQUES

In this study two different kick techniques are distinguished - the direct kick (Fig. 1) and the round kick (Fig. 2). In professional defense, these kicks are used to stop and keep the attacker in distance where the

attacker cannot touch us. The second way of use is destabilization of attacker.

During the direct kick a sole or a heel are the hit areas. This kick is made directly and by the shortest way to the target. During the round kick an instep together with part of shank are hit areas. The direct kick is considered to be stronger than the round kick.



Figure 1: Direct kick

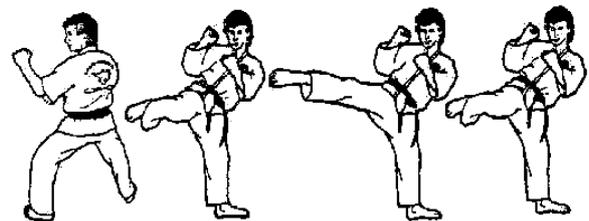


Figure 2: Round kick

MEASURING DEVICES

The strain gauge sensor L6E-C3-300kg (Fig. 3.) works as unilaterally cantilever bending beam. During force delivery the biggest deformation of sensor is in places with the thinnest walls – there are metal film strain gauges which change their electrical resistance depending on deformation. Strain gauges are plugged in Wheatstone bridge and this way is possible to convert difference of resistance to electrical signal which we can process.



Figure 3: Strain gauge sensor L6E-C3-300kg

The sensor is connected to the computer, which is used for data storage, through the strain gauge. The strain gauge type TENZ2334 is an electronic appliance that converts the signals to data that is stored in memory. The core of the appliance is a single-chip microcomputer that controls all of the activities. The strain gauge sensor is connected to this appliance via four-pole connector XLR by four conductors. The number of values measured by the sensor averages around 600 measurements per second while the data is immediately stored in the memory of a device with a capacity of 512 kB (Lapkova et al., 2012).

The mentioned above strain gauge sensor was placed on the measuring station according to the following schematic (Fig. 4):

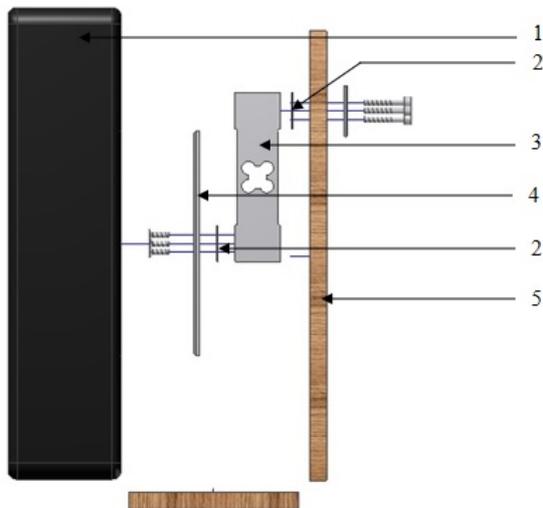


Figure 4: Measuring station schematic

- 1 – punching bag (made from hardened vinyl filled with foam)
- 2 – template
- 3 – strain gauge sensor L6E-C3-300kg
- 4 – board (200 x 200 x 5 mm)
- 5 – punching bag base

EXPERIMENT SETUP

The total of 103 participants took part in the experiment; men and women. All participants were in the age from 19 to 28. Based on previous training and experience the participants were divided into following groups:

- No training – These persons have never done any combat sport, martial art or combat system. They have no theoretical knowledge of the striking technique. The technique was presented to these persons before the experiment for safety reasons. Noted further as M1 (for men) and W1 (for women).
- Mid-trained - These persons have the theoretical knowledge of striking techniques and do attend the Special physical training course for at least

six months. The course is focused on self-defense and professional defense. Noted further as M2 (only men).

- Person who play football are in separate category due to their too specific kicking technique. Noted further as M3 (only men).

The exact numbers of participants in each group are given in Table 1.

Table 1: The number of participants in groups

Group	Number of participants
M1	44
M2	32
M3	18
W1	9

Due to the number of participants in each category the research was focused on male participants. The W1 group was used only for basic comparison between genders.

During the measurement the target was positioned in such manner that the center of the tensometric sensor was in the height of 70cm. The person was made to stay at the same place for the whole experiment. Any unnecessary movement (e. g. lunge etc.) would lead to data distortion.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are inspired in the biological neural nets and are used for complex and difficult tasks (Hertz et al., 1991), (Wasserman, 1980), (Gurney, 1997), (Fausset, 2003). The most often usage is classification of objects as also in this case. ANNs are capable of generalization and hence the classification is natural for them. Some other possibilities are in pattern recognition, control, filtering of signals and also data approximation and others.

There are several kinds of ANN. Simulations were performed with feedforward net with supervision and Levenberg-Marquardt training algorithm (Fausset, 2003). ANN needs a training set of known solutions to be learned on them. Supervised ANN has to have input and also required output. ANN with unsupervised learning exists and there a capability of selforganization is applied.

The neural network works so that suitable inputs in numbers have to be given on the input vector. These inputs are multiplied by weights which are adjusted during the training. In the neuron the sum of inputs multiplied by weights are transferred through mathematical function like sigmoid, linear, hyperbolic tangent etc. to the output from a neuron unit - node.

These single nodes (Fig. 5) are connected to different structures to obtain different structures of ANN (e.g. Fig. 6 and Fig. 7), where $\delta = TF[\sum(w_i x_i + b w_b)]$ and

$\sum = TF[\sum(w_i x_i + b w_b)]$; TF means transfer function and logistic sigmoid function is used in this case.

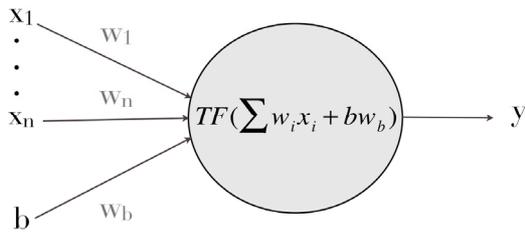


Figure 5: A node model, where TF (transfer function like sigmoid), $x_1 - x_n$ (inputs to neural network), b – bias (usually equal to 1), $w_1 - w_n, w_b$ – weights, y – output

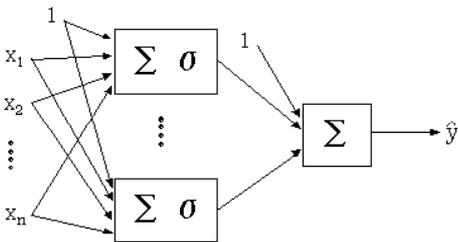


Figure 6: ANN models with one hidden layer

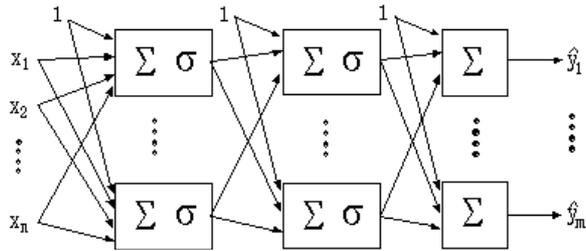


Figure 7: ANN models with two hidden layers and more outputs

The example of relation between inputs and output can be described as a mathematical form (1). It represents the case of only one node and logistic sigmoid function as a transfer function.

$$y = \frac{1}{1 + e^{-(x_1 w_1 + x_2 w_2)}} \quad (1)$$

where: y – output
 x_1, x_2 – inputs
 w_1, w_2 – weights.

PROBLEM DEFINITION AND ANALYSIS

During the experiments on the measuring station ten force profiles for each participant and each kick technique were collected. As an example the mean force profiles for the W1 group are depicted in Fig. 8. The similarities of the main peak are clearly visible. However the mean value may prove very misleading.

In Fig. 9 all collected force profiles for the Round kick (W1 group) are depicted. There is high variety in the force profiles that makes the possibility of simple classification much harder. To improve the chance of successful classification a basic signal processing was applied.

A set of statistical values was used to represent each force profile for the classification. Three different spectral sequences were derived from the force profiles. The first was in the range from 3N to 53N with the bandwidth 10N. The second was in range from 73N to 133N with the bandwidth 20N. Finally the third starting at 201N and ending at 801N with the bandwidth of 200N. By this approach eleven integer number inputs for classification were obtained for each force profile. As last (twelfth) input the rounded median value was used. Mean values of these twelve inputs for W1 group are depicted in figure 10. The aim was to highlight the differences in the signals of different kick techniques.

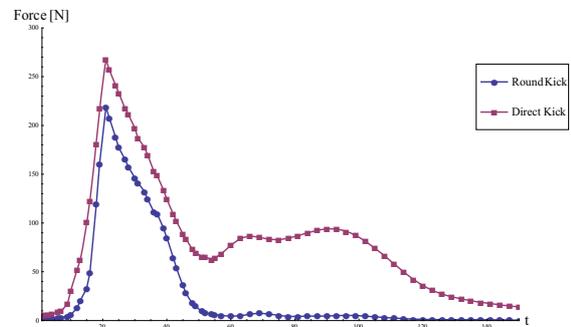


Figure 8: Mean force profiles – Group W1

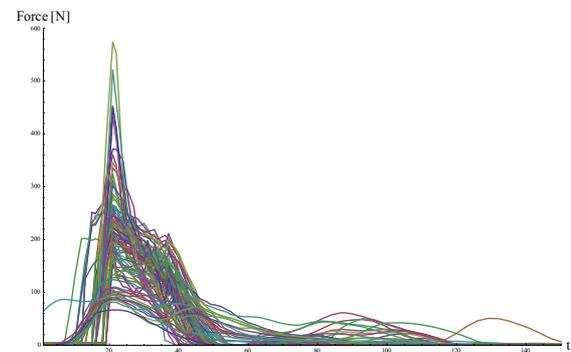


Figure 9: Force profiles – Round kick - Group W1.

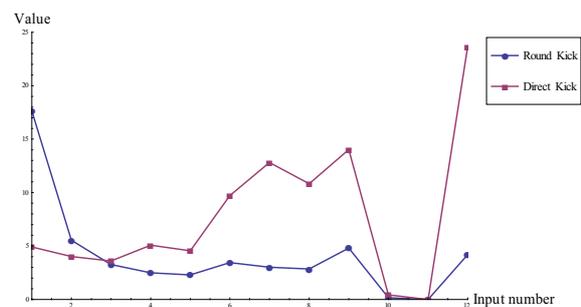


Figure 10: Mean classification input values - Group W1

The example of input values for all force profile samples in group W1 is given in Fig. 11 (Round kick) and in Fig. 12 (Direct kick).

As mentioned previously the simulations were performed with feedforward net with supervision and Levenberg-Marquardt training algorithm. Two different methods of preparing the training and testing set were applied. Typically the set of samples was halved for this purpose. One half was used as a training set. The other half served as testing set. In the first approach (noted “a”) all ten samples for half of the participants in the group were used as the training set. The remaining samples were used for testing set. In the second approach (note “b”) five samples for each participant served as training set and other five as the testing set.

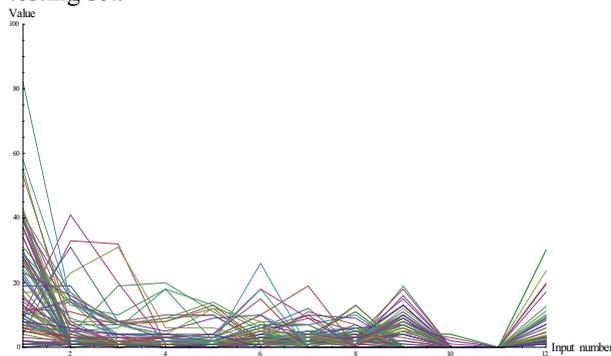


Figure 11: Classification input values – Round kick – Group W1

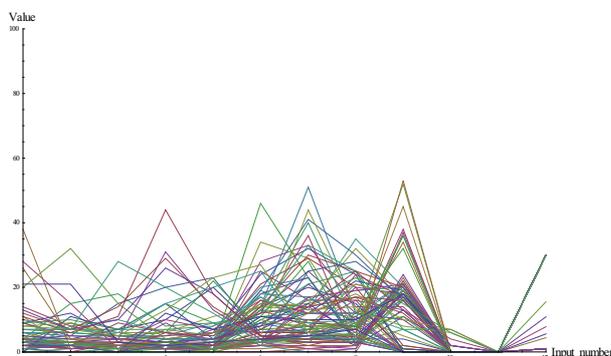


Figure 12: Classification input values – Direct kick – Group W1

Different settings of the number of iterations and number of neurons in the hidden layer were tested. The goal was a higher than 85% rate of successfully classified samples from the testing set. In this initial study the 15% fail rate was taken as acceptable mainly due to errors that occur during the physical measuring. However the aim for future research is to improve the success rate significantly (up to 95% if possible). The best results obtained for each group are presented in following section.

RESULTS

In this section, the results of neural network based classification of different kick techniques are presented. In Table 2 the final neural network setting

for each training set is given alongside with root mean square error (RMSE) that is a typical measure of the quality of training process.

Table 2: Final neural network setting and RMSE

Group	RMSE	Nodes in hidden layer	Iterations
M1a	0.197889	3	60
M1b	0.202538	3	60
M2a	0.172885	2	60
M2b	0.189389	3	20
M3a	2.02E-15	3	40
M3b	0.127763	2	40
W1a	0.255196	7	60
W1b	0.246739	2	30

Furthermore an example of the RMSE shape during the training is given in Fig. 13 for the W1a training set.

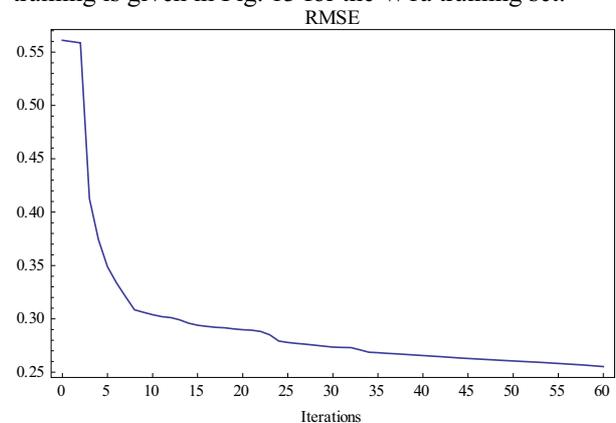


Figure 13: The RMSE shape – W1a training set

In Table 3 the numbers of successfully classified samples (corresponding to results in Table 2) are given.

Table 3: The number of successfully classified samples

Group	N_1/N_2	RMSE	N_3	$N_3 \%$	N_4	$N_4 \%$
M1a	440	0.197	422	95.909	399	90.681
M1b	440	0.202	418	95	407	92.5
M2a	320	0.172	310	96.875	287	89.687
M2b	320	0.189	308	96.25	285	89.062
M3a	180	2.02E-15	180	100	165	91.666
M3b	180	0.127	177	98.333	170	94.444
W1a	80 / 100	0.255	76	95	91	91
W1b	90	0.246	83	92.222	74	82.222

Where:

N_1 – number of samples in the training set.

N_2 – number of samples in the testing set.

Note: Except W1a it applies that $N_1=N_2$.

N_3 – number of successfully classified samples from the training set.

N_4 – number of successfully classified samples from the testing set.

As can be visible from the Table 3, the results achieved at least 89 % which is higher than the specified goal.

CONCLUSION

In this paper an artificial neural network was designed and tested in the task of different kick techniques force profile classification. The goal of 85% successful classification rate was accomplished (in male categories). There have been however several difficulties during the training process. The first was the data preparation issue that was described and solved by employing the spectral analysis. Secondly, the neural network exhibited very strong tendency to over fit, meaning that the higher success rate for training set lead to significantly worse success rate for the testing set. The finding of the balance between these two was the main issue during the neural network designing and learning process. The results presented in previous section are very promising and encourage further research of neural network based classification of force profiles. In the following research the neural networks will be tested on more complex tasks e.g. the classification of participant's gender, training level or for classification of higher number of different kick and striking techniques.

ACKNOWLEDGEMENT

This work was supported by the Internal Grant Agency at TBU in Zlín, project No. IGA/FAI/2014/036,

AUTHOR BIOGRAPHIES

DORA LAPKOVA was born in the Czech Republic, and went to the Tomas Bata University in Zlín, where she studied Security Technologies, Systems and Management and obtained her MSc degree in 2009. She is now a Ph.D. student at the same university. Her email address is: dlapkova@fai.utb.cz.



MICHAL PLUHACEK was born in the Czech Republic, and went to the Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011. He is now a Ph.D. student at the same university. His email address is: pluhacek@fai.utb.cz.



IGA/FAI/2014/10 and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

REFERENCES

- Fausett L. V.: *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*, Prentice Hall, 1993, ISBN: 9780133341867
- Gurney K.: *An Introduction to Neural Networks*, CRC Press, 1997, ISBN: 1857285034
- Hertz J., Kogh A. and Palmer R. G.: *Introduction to the Theory of Neural Computation*, Addison – Wesley 1991
- Lapkova D., Pospisilik M., Adamek M. and Malanik Z.: The utilisation of an impulse of force in self-defence. In: *XX IMEKO World Congress: Metrology for Green Growth*. Busan, Republic of Korea, 2012, ISBN: 978-89-950000-5-2
- Liu, P., et al. A kinematic analysis of round kick in Taekwondo. *ISBS-Conference Proceedings Archive*. Vol. 1. No. 1. 2000.
- Pieter, F., and Pieter W.. Speed and force in selected taekwondo techniques. *Biology of sport* 12, 1995, 257-266.
- Wasserman P. D.: *Neural Computing: Theory and Practice*, Coriolis Group, 1980, ISBN: 0442207433

PSEUDO NEURAL NETWORKS FOR IRIS DATA CLASSIFICATION

Zuzana Kominkova Oplatkova, Roman Senkerik, Ales Kominek

Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{oplatkova, senkerik, kominek}@fai.utb.cz

KEYWORDS

Iris data, Pseudo Neural Network, Analytic programming, Differential Evolution.

ABSTRACT

This research deals with a novel approach to classification. Iris data was used for the experiments. Classical artificial neural networks, where a relation between inputs and outputs is based on the mathematical transfer functions and optimized numerical weights, was an inspiration for this work. Artificial neural networks need to optimize weights, but the structure and transfer functions are usually set up before the training. The proposed method utilizes the symbolic regression for synthesis of a whole structure, i.e. the relation between inputs and output(s). This paper differs from the previous approach where only one output pseudo node was used even for more classes. In this case, there were synthesized more node output equations as in classical artificial neural networks. The benchmark was iris data as in previous research. For experimentation, Differential Evolution (DE) for the main procedure and also for meta-evolution version of analytic programming (AP) was used.

INTRODUCTION

The interest about classification by means of some automatic process has been enlarged with the development of artificial neural networks (ANN). They can be used also for a lot of other possible applications like pattern recognition, prediction, control, signal filtering, approximation, etc. All artificial neural networks are based on some relation between inputs and output(s), which utilizes mathematical transfer functions and optimized weights from training process. The setting-up of layers, number of neurons in layers, estimating of suitable values of weights is a demanding procedure. On account of this fact, pseudo neural networks, which represent the novelty approach using symbolic regression with evolutionary computation, is proposed in this paper.

Symbolic regression in the context of evolutionary computation means to build a complex formula from basic operators defined by users. The basic case represents a process in which the measured data is fitted and a suitable mathematical formula is obtained in an analytical way. This process is widely known for mathematicians. They use this process when a need

arises for mathematical model of unknown data, i.e. relation between input and output values. The symbolic regression can be used also for design of electronic circuits or optimal trajectory for robots and within other applications (Back et al., 1997), (Koza, 1998), (Koza, 1999), (O'Neill et al., 2003), (Zelinka et al., 2011), (Oplatkova, 2009), (Varacha et al., 2006). Everything depends on the user-defined set of operators. The proposed technique is similar to synthesis of analytical form of mathematical model between input and output(s) in training set used in neural networks. Therefore it is called Pseudo Neural Networks.

Initially, John Koza proposed the idea of symbolic regression done by means of a computer in Genetic Programming (GP) (Back et al., 1997), (Koza, 1998), (Koza, 1999). The other approaches are e.g. Grammatical Evolution (GE) developed by Conor Ryan (O'Neill et al., 2003) and here described Analytic Programming (Zelinka et al., 2011), (Oplatkova, 2009), (Varacha et al., 2006).

The above-described tools were recently commonly used for synthesis of artificial neural networks but in a different manner than is presented here. One possibility is the usage of evolutionary algorithms for optimization of weights to obtain the ANN training process with a small or no training error result. Some other approaches represent the special ways of encoding the structure of the ANN either into the individuals of evolutionary algorithms or into the tools like Genetic Programming. But all of these methods are still working with the classical terminology and separation of ANN to neurons and their transfer functions (Fekiac, 2011). In this paper, the proposed technique synthesizes the structure without a prior knowledge of transfer functions and inner potentials. It synthesizes the relation between inputs and output of training set items used in neural networks so that the items of each group are correctly classified according the rules for cost function value. The previous research used a continuous version of classification when just one node is enough even for more classes. This can be done through defined intervals for appropriate group of data. This approach seemed to be easy at the very beginning of this research. On the other hand, the pseudo neural networks should work similarly to artificial neural networks. Therefore the need of more output nodes arises. The combination of output values gives more combinations for number of classes. The case presented in this paper uses three output nodes where each node will encode just one class, i.e. the activated node (output value =1) stands for the

appropriate class, the rest of nodes should be deactivated (the output value = 0). The data set used for training is Iris data set (Machine Learning Repository, Fisher 1936). It is a very known benchmark data set for classification problem, which was introduced by Fisher for the first time.

Firstly, Analytic Programming used as a symbolic regression tool is described. Subsequently Differential Evolution used for main optimization procedure within Analytic Programming and also as a second algorithm within metaevolution purposes is mentioned. After that a brief description of artificial neural network (ANN) follows. Afterwards, the proposed experiment with differences compared to classical ANN is explained. The result section and conclusion finish the paper.

ANALYTIC PROGRAMMING

Basic principles of the AP were developed in 2001 (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova, 2009), (Zelinka et al., 2011). Until that time only genetic programming (GP) and grammatical evolution (GE) had existed. GP uses genetic algorithms while AP can be used with any evolutionary algorithm, independently on individual representation. To avoid any confusion, based on use of names according to the used algorithm, the name - Analytic Programming was chosen, since AP represents synthesis of analytical solution by means of evolutionary algorithms.

The core of AP is based on a special set of mathematical objects and operations. The set of mathematical objects is set of functions, operators and so-called terminals (as well as in GP), which are usually constants or independent variables. This set of variables is usually mixed together and consists of functions with different number of arguments. Because of a variability of the content of this set, it is called here “general functional set” – GFS. The structure of GFS is created by subsets of functions according to the number of their arguments. For example GFS_{all} is a set of all functions, operators and terminals, GFS_{3arg} is a subset containing functions with only three arguments, GFS_{0arg} represents only terminals, etc. The subset structure presence in GFS is vitally important for AP. It is used to avoid synthesis of pathological programs, i.e. programs containing functions without arguments, etc. The content of GFS is dependent only on the user. Various functions and terminals can be mixed together (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova, 2009).

The second part of the AP core is a sequence of mathematical operations, which are used for the program synthesis. These operations are used to transform an individual of a population into a suitable program. Mathematically stated, it is a mapping from an individual domain into a program domain. This mapping consists of two main parts. The first part is called discrete set handling (DSH) (See Figure 1) (Zelinka et al., 2005), (Lampinen and Zelinka, 1999) and the second one stands for security procedures which do not allow synthesizing pathological programs. The method of DSH, when used, allows handling arbitrary objects including nonnumeric objects like linguistic

terms {hot, cold, dark...}, logic terms (True, False) or other user defined functions. In the AP DSH is used to map an individual into GFS and together with security procedures creates the above mentioned mapping which transforms arbitrary individual into a program.

AP needs some evolutionary algorithm (Zelinka, 2004) that consists of population of individuals for its run. Individuals in the population consist of integer parameters, i.e. an individual is an integer index pointing into GFS. The creation of the program can be schematically observed in Fig. 2. The individual contains numbers which are indices into GFS. The detailed description is represented in (Zelinka et al., 2005), (Zelinka et al., 2008), (Oplatkova et al., 2009).

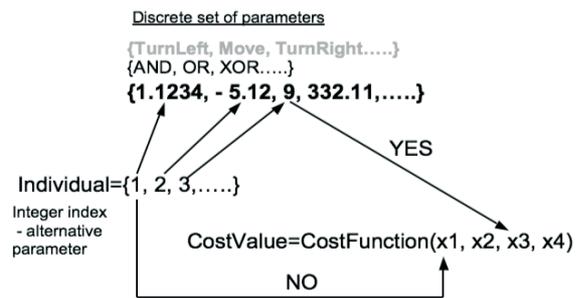
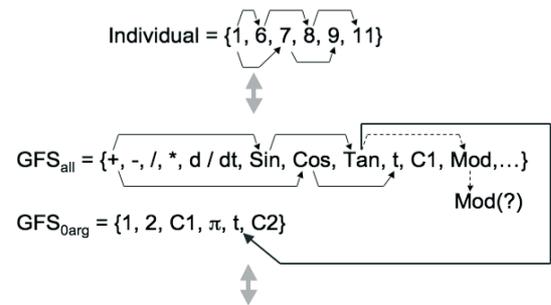


Figure 1: Discrete set handling



Resulting Function by AP = Sin(Tan(t)) + Cos(t)

Figure 2: Main principles of AP

AP exists in 3 versions – basic without constant estimation, AP_{nf} – estimation by means of nonlinear fitting package in Mathematica environment and AP_{meta} – constant estimation by means of another evolutionary algorithms; meta means metaevolution.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are inspired in the biological neural nets and are used for complex and difficult tasks (Hertz et al., 1991), (Wasserman, 1980), (Gurney, 1997), (Fausset, 2003), (Volna et al., 2013). The most often usage is classification of objects as also in this case. ANNs are capable of generalization and hence the classification is natural for them. Some other possibilities are in pattern recognition, control, filtering of signals and also data approximation and others. There are several kinds of ANN. Simulations were based on similarity with feedforward net with

supervision. ANN needs a training set of known solutions to be learned on them. Supervised ANN has to have input and also required output.

The neural network works so that suitable inputs in numbers have to be given on the input vector. These inputs are multiplied by weights which are adjusted during the training. In the neuron the sum of inputs multiplied by weights are transferred through mathematical function like sigmoid, linear, hyperbolic tangent etc. Therefore ANN can be used for data approximation (Hertz et al., 1991) – a regression model on measured data, relation between input and required (measured data) output.

These single neuron units (Fig. 3) are connected to different structures to obtain different structures of ANN (e.g. Fig. 4 and Fig. 5), where $\sum \delta = TF[\sum (w_i x_i + b w_b)]$ and $\sum = TF[\sum (w_i x_i + b w_b)]$; TF is logistic sigmoid function in this case.

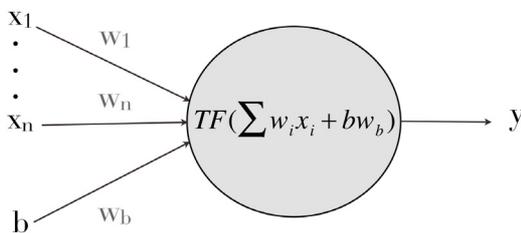


Figure 3: Neuron model, where TF (transfer function like sigmoid), $x_1 - x_n$ (inputs to neural network), b – bias (usually equal to 1), $w_1 - w_n, w_b$ – weights, y – output

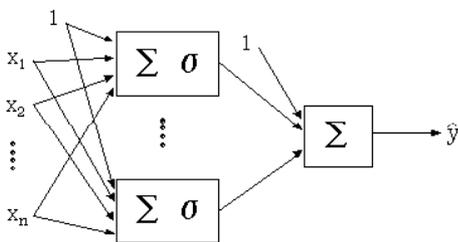


Figure 4: ANN models with one hidden layer

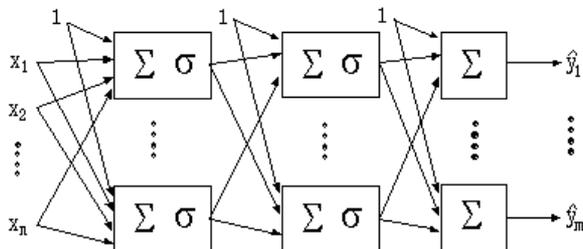


Figure 5: ANN models with two hidden layers and more outputs

The example of relation between inputs and output can be shown as a mathematical form (1). It represents the case of only one neuron and logistic sigmoid function as a transfer function.

$$y = \frac{1}{1 + e^{-(x_1 w_1 + x_2 w_2)}}, \quad (1)$$

where y – output

x_1, x_2 – inputs
 w_1, w_2 – weights.

The aim of the proposed technique is to find similar relation to (1). This relation is completely synthesized by evolutionary symbolic regression – Analytic Programming.

USED EVOLUTIONARY ALGORITHMS

This research used one evolutionary algorithm Differential Evolution (Price, 2005) for both parts – main procedure and metaevolutionary part of the analytic programming. Future simulations expect a usage of soft computing GAHC algorithm (modification of HC12) (Matousek, 2007) and a CUDA implementation of HC12 algorithm (Matousek, 2010).

Differential evolution

DE is a population-based optimization method that works on real-number-coded individuals (Price, 2005). For each individual $\vec{x}_{i,G}$ in the current generation G , DE generates a new trial individual $\vec{x}'_{i,G}$ by adding the weighted difference between two randomly selected individuals $\vec{x}_{r1,G}$ and $\vec{x}_{r2,G}$ to a randomly selected third individual $\vec{x}_{r3,G}$. The resulting individual $\vec{x}'_{i,G}$ is crossed-over with the original individual $\vec{x}_{i,G}$. The fitness of the resulting individual, referred to as a perturbed vector $\vec{u}_{i,G+1}$, is then compared with the fitness of $\vec{x}_{i,G}$. If the fitness of $\vec{u}_{i,G+1}$ is greater than the fitness of $\vec{x}_{i,G}$, then $\vec{x}_{i,G}$ is replaced with $\vec{u}_{i,G+1}$; otherwise, $\vec{x}_{i,G}$ remains in the population as $\vec{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Description of used DERand1Bin strategy is presented in (2). Please refer to (Price and Storn 2001, Price 2005) for the description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \quad (2)$$

PROBLEM DESIGN – IRIS PLANT DATA SET DEFINITION

For this classification problem, iris data set was used (Machine Learning Repository, Fisher 1936). This set contains 150 instances. Half amount was used as training data and the second half was used as testing data. The data set contains 3 classes of 50 instances

each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. Each instance has 4 attributes (sepal length, sepal width, petal length and petal width (Fig. 6)) and type of class – iris virginica (Fig. 7), iris versicolor (Fig. 8) and iris setosa (Fig. 9). The attributes were of real values.



Figure 6: Iris - petal and sepal



Figure 7: Iris virginica



Figure 8: Iris versicolor



Figure 9: Iris setosa

Usually, the class is defined by 3 output nodes in classical artificial neural net and binary code. After the ANN training, it can be separated an equation for each output node - relation between inputs, weights and the output node. The training is done in a parallel way - all weights, which are set up produce the final equations in “one step”. In this paper, AP was able to create pseudo neural net structure between inputs and also three outputs. The synthesis of each relation between inputs and outputs was done in serial and independent way. It is different from the previous approach where only one output node was used and continuous classification within defined intervals of output values for each class. Both approaches seem to be useful. Within more classes, more output nodes are needed. The continuous classification will not be able to cover e.g. 20 classes. There will be very small differences between intervals and it is not usable. The output values were designed as it is usual with ANN in this case. The combination of zeros and ones, i.e. iris setosa was (1,0,0), iris virginica (0,1,0) and iris versicolor (0,0,1).

As the approach is to synthesize three independent output equations, the cost function value in eq. (3) was given the same for all synthesized solutions, i.e. if the cv is equal to zero, all training patterns are classified correctly.

$$cv = \sum_{i=1}^n |requiredOutput - currentOutput| \quad (3)$$

The training patterns were setup as the appropriate item in the correct class to 1 and the other two kinds of plants were zero and this was cyclically done for all cases.

RESULTS

As described in section about Analytic Programming, AP requires some EA for its run. In this paper AP_{meta} version was used. Meta-evolutionary approach means usage of one main evolutionary algorithm for AP process and second algorithm for coefficient estimation, thus to find optimal values of constants in the structure of pseudo neural networks.

In this paper, DE was used for main AP process and also in the second evolutionary process. Settings of EA parameters for both processes were based on performed numerous experiments with chaotic systems and simulations with AP_{meta} (Table 1 and Table 2).

Table 1: DE settings for main process

PopSize	20
F	0.8
CR	0.8
Generations	50
Max. CF Evaluations (CFE)	1000

Table 2: DE settings for meta-evolution

PopSize	40
F	0.8
CR	0.8
Generations	150
Max. CF Evaluations (CFE)	6000

REFERENCES

- Back T., Fogel D. B., Michalewicz Z., *Handbook of evolutionary algorithms*, Oxford University Press, 1997, ISBN 0750303921
- Fausett L. V.: *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*, Prentice Hall, 1993, ISBN: 9780133341867
- Fekiac J., Zelinka I., Burguillo J. C.: A review of methods for encoding neural network topologies in evolutionary computation, ECMS 2011, Krakow, Poland, ISBN: 978-0-9564944-3-6
- Fisher R. A. (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics* 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x
- Gurney K.: *An Introduction to Neural Networks*, CRC Press, 1997, ISBN: 1857285034
- Hertz J., Kogh A. and Palmer R. G.: *Introduction to the Theory of Neural Computation*, Addison – Wesley 1991
- Koza J. R. et al., *Genetic Programming III; Darwinian Invention and problem Solving*, Morgan Kaufmann Publisher, 1999, ISBN 1-55860-543-6
- Koza J. R., *Genetic Programming*, MIT Press, 1998, ISBN 0-262-11189-6
- Lampinen J., Zelinka I., 1999, "New Ideas in Optimization – Mechanical Engineering Design Optimization by Differential Evolution", Volume 1, London: McGraw-hill, 1999, 20 p., ISBN 007-709506-5.
- Machine learning repository with Iris data set <http://archive.ics.uci.edu/ml/datasets/Iris>
- Matousek R., 2010, „HC12: The Principle of CUDA Implementation“. In MENDEL 2010, Mendel Journal series, pp. 303-308. ISBN: 978-80-214-4120- 0. ISSN: 1803- 3814.
- Matousek R., 2007, „GAHC: Improved GA with HC station“, In WCECS 2007, San Francisco, pp. 915-920. ISBN: 978-988-98671-6-4.
- O’Neill M., Ryan C., *Grammatical Evolution. Evolutionary Automatic Programming in an Arbitrary Language*, Kluwer Academic Publishers, 2003, ISBN 1402074441
- Oplatkova Z.: *Metaevolution: Synthesis of Optimization Algorithms by means of Symbolic Regression and Evolutionary Algorithms*, Lambert Academic Publishing Saarbrücken, 2009, ISBN: 978-3-8383-1808-0
- Price K., Storn R. M., Lampinen J. A., 2005, "Differential Evolution : A Practical Approach to Global Optimization", (Natural Computing Series), Springer; 1 edition.
- Price, K. and Storn, R. (2001), *Differential evolution homepage*, [Online]: <http://www.icsi.berkeley.edu/~storn/code.html>, [Accessed 29/02/2012].
- Volna, E. Kotyrba, M. and Jarusek, R. 2013 "Multiclassifier based on Elliott wave’s recognition" *Computers and Mathematics with Applications* 66 (2013) ISSN: 0898-1221.DOI information: 10.1016/j.camwa.2013.01.012.
- Wasserman P. D.: *Neural Computing: Theory and Practice*, Coriolis Group, 1980, ISBN: 0442207433
- Zelinka et al.: *Analytical Programming - a Novel Approach for Evolutionary Synthesis of Symbolic Structures*, in Kita E.: *Evolutionary Algorithms*, InTech 2011, ISBN: 978-953-307-171-8
- Zelinka I., Varacha P., Oplatkova Z., *Evolutionary Synthesis of Neural Network*, Mendel 2006 – 12th International Conference on Softcomputing, Brno, Czech Republic, 31 May – 2 June 2006, pages 25 – 31, ISBN 80-214-3195-4
- Zelinka I., Oplatkova Z., Nolle L., 2005. *Boolean Symmetry Function Synthesis by Means of Arbitrary Evolutionary Algorithms-Comparative Study*, *International Journal of Simulation Systems, Science and Technology*, Volume 6, Number 9, August 2005, pages 44 - 56, ISSN: 1473-8031.

AUTHOR BIOGRAPHIES

ZUZANA KOMINKOVA OPLATKOVA is an associate professor at Tomas Bata University in Zlin.



Her research interests include artificial intelligence, soft computing, evolutionary techniques, symbolic regression, neural networks. She is an author of around 100 papers in journals, book chapters and conference proceedings. Her e-mail address

is: oplatkova@fai.utb.cz

ROMAN SENKERIK is an associate professor at Tomas Bata University in Zlin. His research interests include artificial intelligence, soft computing,



evolutionary techniques, theory of deterministic chaos. He is an author of more than 100 papers in journals, book chapters and conference proceedings. His email address is: senkerik@fai.utb.cz

ALES KOMINEK was born in the Czech Republic, and went to the Tomas Bata University in Zlin, where he studied Technical Cybernetics and obtained his MSc degree in 2002. He has been studying Ph.D. degree in Engineering Informatics currently since 2012. His email address is: kominek@fai.utb.cz



SIMULATION OF THE DIFFERENTIAL EVOLUTION PERFORMANCE DEPENDENCY ON SWITCHING OF THE DRIVING CHAOTIC SYSTEMS

¹Roman Senkerik, ¹Michal Pluhacek, ²Donald Davendra, ²Ivan Zelinka, ¹Zuzana Kominkova Oplatkova

¹Tomas Bata University in Zlin , Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{senkerik , oplatkova , pluhacek}@fai.utb.cz

²Department of Computer Science, Faculty of Electrical Engineering and Computer Science
VB-TUO, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
{donald.davendra , ivan.zelinka}@vsb.cz

KEYWORDS

Deterministic chaos; Discrete chaotic maps; Evolutionary computation; Differential Evolution; Chaotic Pseudo Random Number Generators

ABSTRACT

This research deals with the deeper analysis of the novel concept of a multi-chaos-driven evolutionary algorithm Differential Evolution (DE). This paper is aimed at the embedding and alternating of set of two discrete dissipative chaotic systems in the form of chaos pseudo random number generator for DE. Repeated simulations were performed on the selected test function in higher dimensions. Finally, the obtained results are compared with canonical DE.

INTRODUCTION

These days the methods based on soft computing such as neural networks, evolutionary algorithms, fuzzy logic, and genetic programming are known as powerful tool for almost any difficult and complex optimization problem. Differential Evolution (DE) (Price 1999) is one of the most potent heuristics available.

This paper is aimed at the investigating the novel concept of multi-chaos driven DE. Although a number of DE variants have been recently developed, the focus of this paper is the embedding of chaotic systems in the form of chaos pseudo random number generator (CPRNG) into the DE (ChaosDE).

Firstly, the motivation for this research is proposed. The next sections are focused on the description of evolutionary algorithm DE, the concept of chaos driven DE and the used test function. Results and conclusion follow afterwards.

MOTIVATION

This research is an extension and continuation of the previous successful initial experiments with chaos driven DE (Senkerik et al. 2014), (Senkerik et al. 2013) with test functions in higher dimensions.

In this paper the novel initial concept of DE/rand/1/bin strategy driven alternately by two chaotic maps

(systems) is more deeply studied. From the previous research, it follows that very promising results were obtained through the utilization of different chaotic maps within the ChaosDE concept. The idea was then to connect these several different influences given by different CPRNGs to the performance of DE into the one multi-chaotic concept. This paper is aimed to the deeper analysis of the novel Multi-ChaosDE concept and the performance dependency on switching of the driving chaotic systems.

Recent research in chaos driven heuristics has been fueled with the predisposition that unlike stochastic approaches, a chaotic approach is able to bypass local optima stagnation. A chaotic approach generally uses the chaotic map in the place of a pseudo random number generator (Aydin et al. 2010). This causes the heuristic to map unique regions, since the chaotic map iterates to new regions. The task is then to select a very good chaotic map as the pseudo random number generator.

The initial concept of embedding chaotic dynamics into the evolutionary algorithms is given in (Caponetto et al. 2003). Later, the initial study (Davendra et al. 2010) was focused on the simple embedding of chaotic systems into the DE in the form of chaos pseudo random number generator (CPRNG). Also the PSO (Particle Swarm Optimization) algorithm with elements of chaos was introduced as CPSO (Coelho and Mariani 2009). The chaos embedded PSO with inertia weigh strategy was closely investigated (Pluhacek et al. 2013a) afterwards, followed by the introduction of a PSO strategy driven alternately by two chaotic systems (Pluhacek et al. 2013b). The primary aim of this work is not to develop a new type of pseudo random number generator, which should pass many statistical tests, but to try to use and test the implementation of natural chaotic dynamics into evolutionary algorithm as a multi-chaotic pseudo random number generator.

DIFFERENTIAL EVOLUTION

DE is a population-based optimization method that works on real-number-coded individuals (Price 1999). For each individual $\vec{x}_{i,G}$ in the current generation G, DE generates a new trial individual $\vec{x}'_{i,G}$ by adding the

weighted difference between two randomly selected individuals $\bar{x}_{r1,G}$ and $\bar{x}_{r2,G}$ to a randomly selected third individual $\bar{x}_{r3,G}$. The resulting individual $\bar{x}'_{i,G}$ is crossed-over with the original individual $\bar{x}_{i,G}$. The fitness of the resulting individual, referred to as a perturbed vector $\bar{u}_{i,G+1}$, is then compared with the fitness of $\bar{x}_{i,G}$. If the fitness of $\bar{u}_{i,G+1}$ is greater than the fitness of $\bar{x}_{i,G}$, then $\bar{x}_{i,G}$ is replaced with $\bar{u}_{i,G+1}$; otherwise, $\bar{x}_{i,G}$ remains in the population as $\bar{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Please refer to (Price 1999), (Price et al. 2005) for the detailed description of the used DERand1Bin strategy (1) (both for Chaos DE and Canonical DE) as well as for the complete description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \quad (1)$$

ChaosDE AND MultiChaosDE CONCEPT

The general idea of ChaosDE and CPRNG is to replace the default PRNG with the discrete chaotic map. As the discrete chaotic map is a set of equations with a static start position, we created a random start position of the map, in order to have different start position for different experiments (runs of EA's). This random position is initialized with the default PRNG, as a one-off randomizer. Once the start position of the chaotic map has been obtained, the map generates the next sequence using its current position.

From the previous research it follows, that very promising results were obtained through the utilization of Delayed Logistic, Lozi, Burgers and Tinkerbelt chaotic maps within the (single) ChaosDE concept. The last two mentioned chaotic maps have unique properties with connection to DE: strong progress towards global extreme, but weak overall statistical results, like average CF value and std. dev., and tendency to premature stagnation. While through the utilization of the Lozi and Delayed Logistic map the continuously stable and very satisfactory performance of ChaosDE was achieved. The idea is then to connect these two different influences to the performance of DE into the one multi-chaotic concept (Multi-ChaosDe).

SELECTED DISCRETE CHAOTIC SYSTEMS

This section contains the description of discrete dissipative chaotic maps used as the chaotic pseudo random generators for DE. In this research, direct output iterations of the chaotic maps were used for the generation of real numbers in the process of crossover based on the user defined CR value and for the generation of the integer values used for the selection of individuals. Following chaotic maps were used: Burgers (2), and Lozi map (3).

The typical chaotic behavior of the utilized maps, represented by the examples of direct output iterations is depicted in Fig. 1 (Burgers map) and Fig. 3 (Lozi map). The illustrative histograms of the distribution of real numbers transferred into the range $\langle 0 - 1 \rangle$ generated by means of studied chaotic maps are in Figures 2 and 4.

Burgers Map

The Burgers mapping is a discretization of a pair of coupled differential equations which were used to illustrate the relevance of the concept of bifurcation to the study of hydrodynamics flows. The map equations are given in (2) with control parameters $a = 0.75$ and $b = 1.75$ as suggested in (Sprott 2003).

$$\begin{aligned} X_{n+1} &= aX_n - Y_n^2 \\ Y_{n+1} &= bY_n + X_nY_n \end{aligned} \quad (2)$$

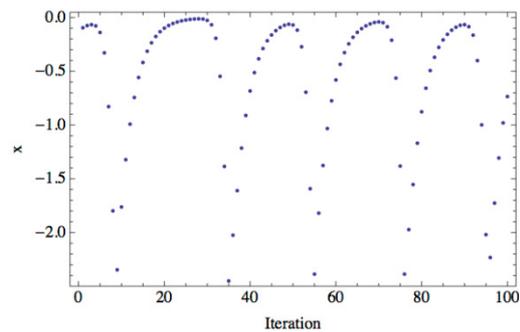


Figure 1: Iterations of the Burgers map (variable x)

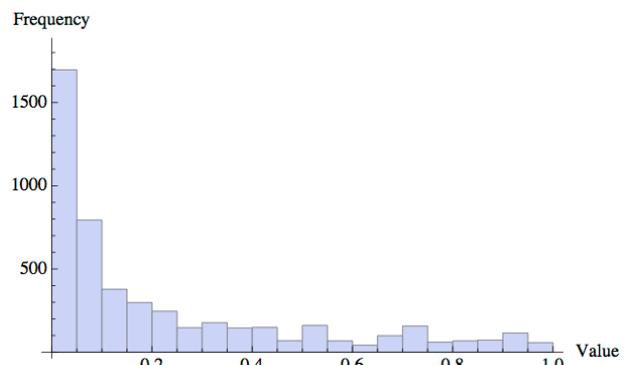


Figure 2: Histogram of the distribution of real numbers transferred into the range $\langle 0 - 1 \rangle$ generated by means of the chaotic Burgers map – 5000 samples

Lozi map

The Lozi map is a discrete two-dimensional chaotic map. The map equations are given in (3). The parameters used in this work are: $a = 1.7$ and $b = 0.5$ as suggested in (Sprott 2003). For these values, the system exhibits typical chaotic behavior and with this parameter setting it is used in the most research papers and other literature sources.

$$\begin{aligned} X_{n+1} &= 1 - a|X_n| + bY_n \\ Y_{n+1} &= X_n \end{aligned} \quad (3)$$

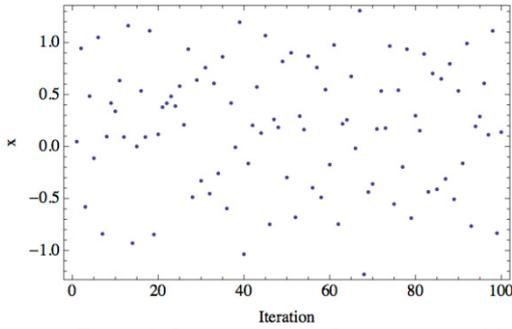


Figure 3: Iterations of the Lozi map (variable x)

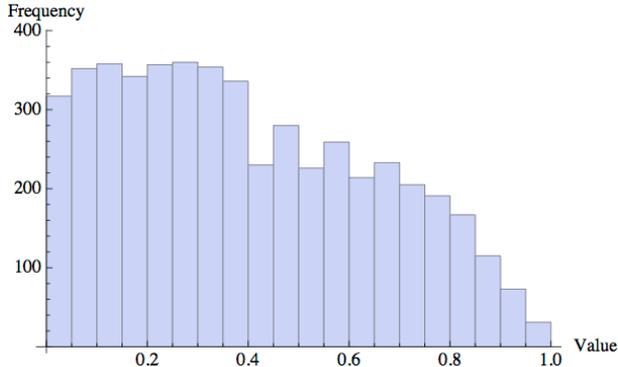


Figure 4: Histogram of the distribution of real numbers transferred into the range $\langle 0 - 1 \rangle$ generated by means of the chaotic Lozi map – 5000 samples

EXPERIMENT DESIGN

For the purpose of evolutionary algorithm performance comparison within this initial research, the multimodal Schwefel's test function (4) was selected.

$$f(x) = \sum_{i=1}^D -x_i \sin(\sqrt{|x_i|}) \quad (4)$$

Function minimum: Position for E_n :

$(x_1, x_2, \dots, x_n) = (420.969, 420.969, \dots, 420.969)$

Value for E_n : $y = -418.983 \cdot Dimension$

The novelty of this research represents the simulation of the DE performance dependency on switching of the driving chaotic systems.

In this paper, the canonical DE strategy DERand1Bin and the Multi-Chaos DERand1Bin strategy driven alternately by two different chaotic maps (ChaosDE) were used.

The parameter settings for both canonical DE and ChaosDE were obtained analytically based on numerous experiments and simulations (see Table 1).

Table 2: DE settings for meta-evolution

DE Parameter	Value
PopSize	75
F	0.8
CR	0.8
Generations	3000
Max. CF Evaluations (CFE)	225000

Investigation on the moment of manual switching over between two chaotic maps represents the main aim of this paper.

Experiments were performed in the combined environment of *Wolfram Mathematica* and *C language*, canonical DE therefore used the built-in *C language* pseudo random number generator *Mersenne Twister C* representing traditional pseudorandom number generators in comparisons. All experiments used different initialization, i.e. different initial population was generated within the each run of Canonical or Chaos driven DE.

EXPERIMENT RESULTS

This initial research utilizes the maximum number of generations fixed at 3000 generations. This allowed the possibility to analyze the progress of DE within a limited number of generations and cost function evaluations.

The statistical results of the experiments are shown in Table 2, which represent the simple statistics for cost function (CF) values, e.g. average, median, maximum values, standard deviations and minimum values representing the best individual solution for all 50 repeated runs of canonical DE and several versions of ChaosDE and Multi-ChaosDE.

Table 3 compares the progress of several versions of ChaosDE, Multi-ChaosDE and Canonical DE. This table contains the average CF values for the generation No. 750, 1500, 2250 and 3000 from all 50 runs. The bold values within the both Tables 2 and 3 depict the best obtained results. Following versions of Multi-ChaosDE were studied:

Burgers-Lozi-Switch-500: Start with Burgers map CPRNG, switch to the Lozi map CPRNG after 500 generations.

Burgers-Lozi-Switch-1500: Start with Burgers map CPRNG, switch to the Lozi map CPRNG after 1500 generations.

Lozi-Burgers-Switch-500: Start with Lozi map CPRNG, switch to the Burgers map CPRNG after 500 generations.

Lozi-Burgers-Switch-1500: Start with Lozi map CPRNG, switch to the Burgers map CPRNG after 1500 generations.

The graphical comparison of the time evolution of average CF values for all 50 runs of ChaosDE/Multi-ChaosDE and canonical DERand1Bin strategy is depicted in Fig. 6. Finally the Figures 5 a) – 5d) confirm the robustness of Multi-ChaosDE in finding the best solutions for all 50 runs.

Obtained numerical results given in Tables 2 and 3 and graphical comparisons in Figures 5 and 6 support the claim that all Multi-Chaos/ChaosDE versions have given better overall results in comparison with the canonical DE version. From the presented data it follows, that Multi-Chaos DE versions driven by Lozi/Burgers Map have given the best overall results.

Table 2: Simple results statistics for the Schwefel's function – 30D

DE Version	Avg CF	Median CF	Max CF	Min CF	StdDev
Canonical DE	-5957.28	-5919.58	-5486.45	-6553.09	272.7228
Burger-Lozi-Switch-500	-11306.1	-11326.5	-9153.31	-12387.9	677.7153
Burger-Lozi-Switch-1500	-10982.9	-11067	-9832.01	-12153.4	530.9785
Lozi-Burger-Switch-500	-11120.7	-11188.4	-9794.39	-12208.5	515.4589
Lozi-Burger-Switch-1500	-11480.5	-11619.6	-10384	-12321.3	479.3151

Table 3: Comparison of progress towards the minimum for the Schwefel's function

DE Version	Generation No.: 750	Generation No.: 1500	Generation No.: 2250	Generation No.: 3000
Canonical DE	-5281.95	-5529.28	-5749.8	-5957.28
Burger-Lozi-Switch-500	-6466.86	-8660.88	-10360.7	-11306.1
Burger-Lozi-Switch-1500	-6845.26	-9916.04	-10738.9	-10982.9
Lozi-Burger-Switch-500	-5957.77	-8692.1	-10680.1	-11120.7
Lozi-Burger-Switch-1500	-5874.04	-7949.73	-10808.9	-11480.5

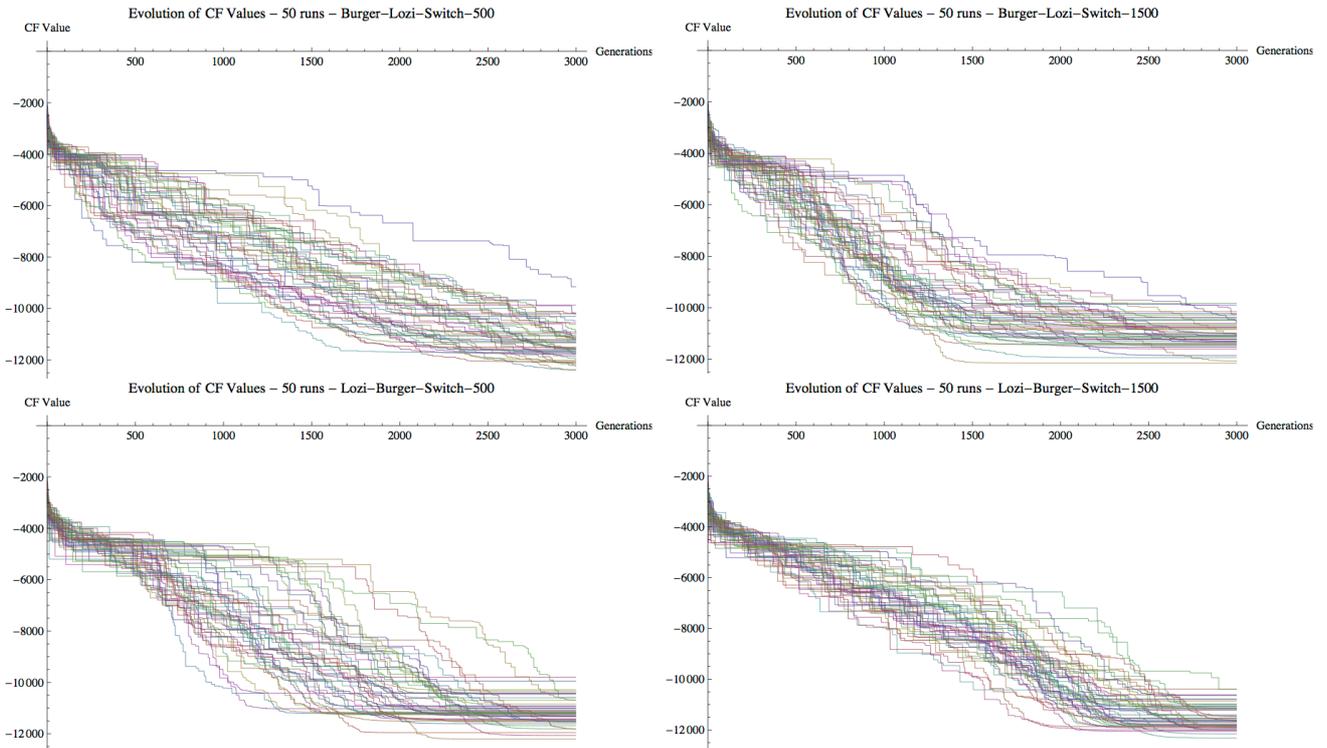


Figure 5: Comparison of the time evolution of CF values for all 50 runs of Multi-ChaosDE version: 5 a) (upper left) Burgers-Lozi-Switch-500; 5 b) (upper right) Burgers-Lozi-Switch-1500; 5 c) (below left) Lozi-Burgers-Switch-500; 5 d) (below right) Lozi-Burgers-Switch-1500

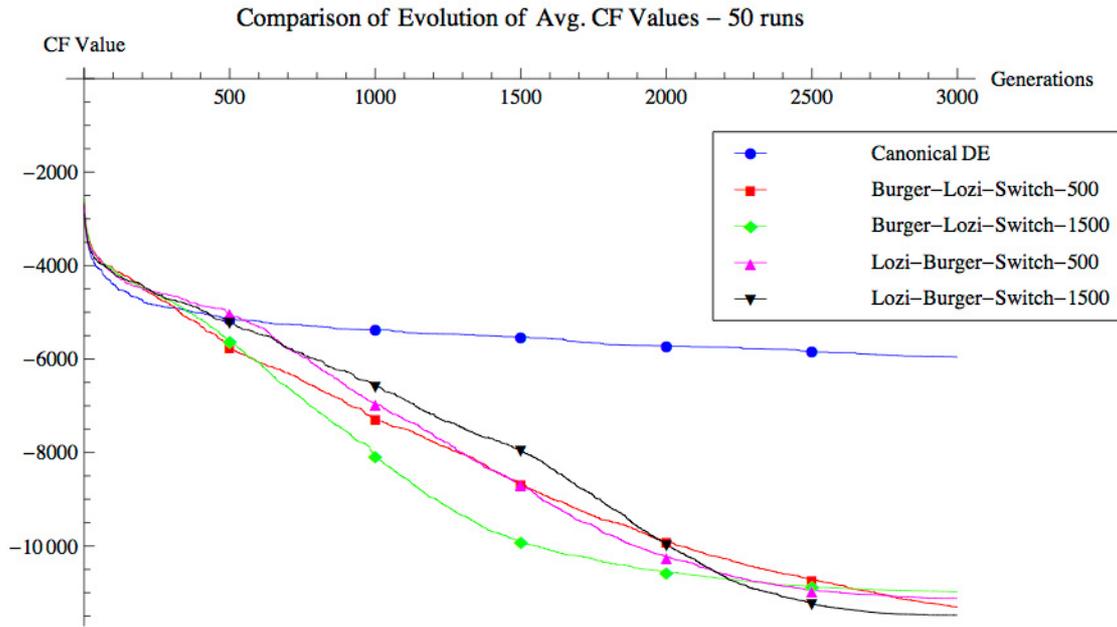


Figure 6: Comparison of the time evolution of avg. CF values for the all 50 runs of Canonical DE, ChaosDE and Multi-ChaosDE. Schwefel's function, $D = 30$.

RESULTS ANALYSIS

For the both *Burgers-Lozi-Switch* versions the progressive Burgers map CPRNG secured the faster approaching towards the global extreme from the very beginning of evolutionary process. The very fast switch over to the Lozi map based CPRNG (*Burgers-Lozi-Switch-500* version) helped to avoid the Burgers map based CPRNG weak spots, which are the weak overall statistical results, like average CF value and std. dev.; and tendency to stagnation. This version was able to reach the best individual minimum CF value. The aforementioned weak spots of the Burgers map based CPRNG have fully revealed in the case of later alternating of both maps. The initial faster convergence (starting of evolutionary process) and subsequent continuously stable searching process without premature stagnation issues are visible from Fig. 6 (red and green lines).

Through the utilization of *Lozi-Burgers-Switch* versions, the strong progress towards global extreme given by Burgers map CPRNG helped to the evolutionary process driven moderately from the start by means of Lozi map CPRNG to achieve the best avg. CF and median CF values. The moment of switch (at 500 and 1500 generations) is clearly visible from Fig. 6 (magenta and black lines). From the results, it seems that it is better to keep the Lozi map based CPRNG for more generations to ensure the stable searching process.

CONCLUSION

In this paper, the novel concept of multi-chaos driven DERand1Bin strategy was more deeply analyzed and compared with the canonical DERand1Bin strategy on the selected benchmark function in higher dimension. Based on obtained results, it may be claimed, that the developed Multi-ChaosDE gives considerably better results than other compared heuristics.

The novelty of this research represents the deeper investigation and simulation of the DE performance dependency on switching of the two driving chaotic systems.

Future plans are including the testing of combination of different chaotic systems as well as the adaptive switching and obtaining a large number of results to perform statistical tests.

Furthermore chaotic systems have additional parameters, which can be tuned. This issue opens up the possibility of examining the impact of these parameters to generation of random numbers, and thus influence on the results obtained by means of ChaosDE.

ACKNOWLEDGEMENT

Grant Agency of the Czech Republic - GACR P103/13/08195S, is partially supported by Grants of SGS No. SP2014/159 and SP2014/170, VŠB - Technical University of Ostrava, Czech Republic, by the Development of human resources in research and development of latest soft computing methods and their application in practice project, reg. no. CZ.1.07/2.3.00/20.0072 funded by Operational

Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic, further was supported by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089 and by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2014/010.

REFERENCES

- Aydin I., Karakose M., Akin E. 2010, "Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection", *Expert Systems with Applications*, Vol. 37, No. 7, pp. 5285–5294.
- Caponetto R., Fortuna L., Fazzino S., and Xibilia M.G. 2010. "Chaotic sequences to improve the performance of evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, pp. 289-304..
- Coelho L.d.S., and Mariani V.C. 2009. "A novel chaotic particle swarm optimization approach using Hénon map and implicit filtering local search for economic load dispatch," *Chaos, Solitons & Fractals*, vol. 39, pp. 510-518.
- Davendra D., Zelinka I., and Senkerik R. 2010. "Chaos driven evolutionary algorithms for the task of PID control," *Computers & Mathematics with Applications*, vol. 60, pp. 1088-1104.
- Pluhacek M., Senkerik R., Zelinka I., and Davendra D. 2013b. "Chaos PSO algorithm driven alternately by two different chaotic maps - An initial study," in *2013 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2444-2449.
- Pluhacek, M., Senkerik, R., Davendra, D., Kominkova Oplatkova, Z., and Zelinka, I. 2013a. "On the behavior and performance of chaos driven PSO algorithm with inertia weight," *Computers & Mathematics with Applications*, vol. 66, pp. 122-134.
- Price K. V. 1999. "An Introduction to Differential Evolution," in *New Ideas in Optimization*, D. Corne, M. Dorigo, and F. Glover, Eds., ed: McGraw-Hill Ltd., pp. 79-108.
- Price, K.V., Storn, R.M., and Lampinen, J.A. 2005. *Differential Evolution - A Practical Approach to Global Optimization*: Springer Berlin Heidelberg.
- Senkerik R., Pluhacek M., Davendra D., Zelinka I., and Kominkova Oplatkova Z. 2013. "Chaos driven evolutionary algorithm: A new approach for evolutionary optimization," *International Journal of Mathematics and Computers in Simulation*, vol. 7, pp. 363-368.
- Senkerik R., Pluhacek M., Zelinka I., Oplatkova Z., Vala R., and Jasek R. 2014. "Performance of Chaos Driven Differential Evolution on Shifted Benchmark Functions Set," in *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. vol. 239, Á. Herrero, B. Baruque, F. Klett, A. Abraham, V. Snášel, A. C. P. L. F. Carvalho, et al., Eds., ed: Springer International Publishing, pp. 41-50.
- Sprott J. C. 2003. *Chaos and Time-Series Analysis*: Oxford University Press.

Modelling and Simulation in Robotic Applications

GYROSCOPIC PRECESSION IN MOTION MODELLING OF BALL-SHAPED ROBOTS

Tomi Ylikorpi
Pekka Forsman
Aarne Halme

Department of Electrical Engineering and Automation
Aalto University
P.O. Box 15500, 00076 Aalto, Finland
E-mail: tomi.ylikorpi@aalto.fi

KEYWORDS

Modelling and simulation in robotic applications, ball-shaped robots, robot dynamics, multi-body simulation

ABSTRACT

This study discusses kinematic and dynamic precession models for a rolling ball with a finite contact area and a point contact respectively. In literature, both conventions have been applied. In this paper, we discuss in detail the kinematic and dynamic models to describe the ball precession and the radius of a circular rolling path. The kinematic model can be used if the contact area and friction coefficient are sufficient to prevent slippage. The dynamic precession model has significance in multi-body simulation environments handling rolling balls with ideal point contacts. We have applied both the kinematic and dynamic precession model to evaluate the no-slip condition of the existing GimBall-robot. According to the result, the necessity of an external precession torque may cause slipping at lower velocities than expected if ignoring this torque.

SYMBOLS

d	Contact area diameter (m)
I_0	Main moment of inertia about rolling axis ($\text{kg}\cdot\text{m}^2$)
I	Two other main moments of inertia ($\text{kg}\cdot\text{m}^2$)
I_{pz}	Pendulum inertia about vertical axis ($\text{kg}\cdot\text{m}^2$)
M_ζ	Forward rolling torque (N·m)
M_ξ	Sideways roll torque (N·m)
M_η	Torque around η -axis (N·m)
r	Path radius (m)
r_c	Path radius, to contact area center (m)
R	Ball radius (m)
R'	Effective rolling radius (m)
t	Time (s)
T_z	Torque around ground vertical Z-axis (N·m)
γ	Forward rolling angle of the ball, pitch angle (rad)
ζ	Forward rolling axis of the ball
ξ	Sideways roll axis of the ball
η	Axis orthogonal to ζ and ξ .
θ	Sideways roll angle of the ball, lean angle (rad)
$\dot{\theta}$	Sideways roll rate (rad/s)
$\ddot{\theta}$	Sideways roll acceleration (rad/s ²)
φ	Heading angle of the ball, yaw angle (rad)

$\Omega, \dot{\varphi}$	Precession rate, spin rate (rad/s)
$\ddot{\varphi}$	Precession acceleration of the ball (rad/s ²)
ω	Forward rolling rate of the ball (rad/s)
$\dot{\omega}$	Forward rolling acceleration (rad/s ²)
μ	Friction coefficient at contact
μ_a	Friction coefficient share for acceleration force
μ_τ	Friction coefficient share for vertical torque

INTRODUCTION

Ball-shaped vehicles have been under development already over the last 120 years. The first patents on self-propelled spherical toys were filed in the end of 19th century. Studies on dynamic modelling and steering of a motor-driven ball started in 1990's leading into emergence of computer controlled spherical mobile robots. (Ylikorpi and Suomela 2007) Recent studies on ball-shaped robots have described a variety of applications in different environments, including marine, indoors, outdoors and planetary exploration. (Michaud and Caron 2002; Bruhn et al. 2005; Kaznov and Seeman 2010) Lately, commercial spherical robots have been introduced to the markets for surveillance and gaming applications. (Avery 2011; Krieger 2013)

Spherical rolling robots offer interesting modelling and control problems due to their extraordinary dynamic nature. In development of robot mechanics and control, simulators regularly represent the robotic system and its behavior. (For example: Hristu-Varsakelis 2001; Otani et al. 2006; Jia et al. 2008; Liu et al. 2008; Ghanbari et al. 2010; Ishikawa et al. 2010; Sang et al. 2011; Zheng et al. 2011; Cai et al. 2012)

Figure 1 presents the pendulum-driven GimBall-robot developed at Aalto University. Steering of this robot takes place by tilting the rolling axis sideways with the aid of the pendulum. As shown in Figure 2, the ball then adopts a circular rolling path. The path center appears to be located at the crossing point of the tilted rolling axis extension and the rolling plane.

Although this behavior appears trivial at the first glance, several questions can be formulated:

1. What makes the ball to follow a circular path?
2. Is the path center actually located at the extension of the rolling axis? If so, why?

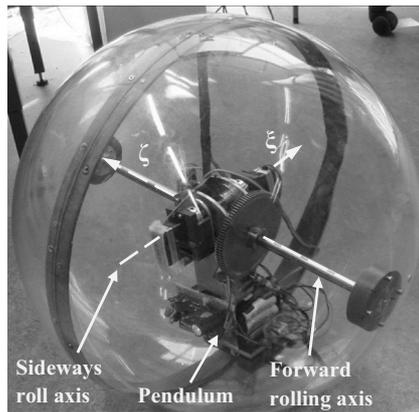


Figure 1: The *GimBall* -robot in Aalto University

3. Assuming no slipping or sliding, can the behavior be explained purely by geometry?
4. When the ball velocity changes, a torque around a vertical axis is needed to change the angular precession-rate. Where this torque comes from?
5. In particular, how behaves an ideal ball on an ideal plane with an ideal point contact? This is a relevant question for simulation models.
6. How do these findings reflect into practice?

This study seeks for answers for the above set questions. In particular, we are looking for the relationship between the lean angle and the path radius. The results find significance in formulation of accurate dynamic models for ball robots, and in application of those models for simulation purposes.

For our robot, forward rolling takes place around an axis whose sideways lean angle with respect to the rolling plane is assumed stable and under control. Forward and backward rolling is unlimited but sideways motion limits to the adjustment of the sideways lean angle only. Steering of the rolling direction to the right or left requires a precession motion of the rolling axis (and the entire robot) around the vertical axis. Omni-directional rolling, that can take place around any axis at any moment, is not possible for this kind of pendulum-driven robots.

RELATED WORK

Literature presents basically two different models for ball precession: it is described either as a kinematic behavior or as a dynamic one. The convention selected in the literature follows in general the mechanical design of the robot being discussed. It is in our interest to find out, whether the kinematic or a dynamic model should be used for a pendulum-driven robot, and under which circumstances is the selection valid. This section gives an overview on the ball precession models in the literature. The discussion is divided into five subsections, each considering a specific approach.

Classical ball-plate problem, spinning forbidden

Prior studies considering the rolling contact between

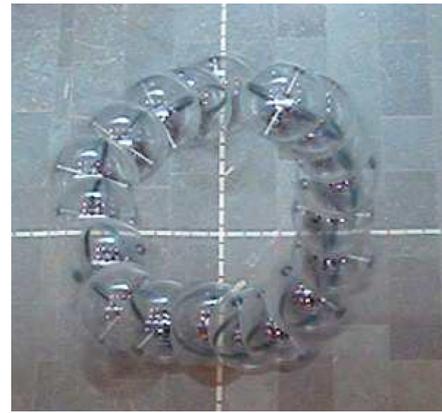


Figure 2: GimBall Follows a Circular Trajectory.
Adopted from Nagai (2008)

two rigid bodies with curved surfaces lay ground for further path planning and control of rolling balls. One subset of the rolling body problem is a sphere rolling along a plane. In this set-up, the ball moves along a planar surface through rotations around horizontal axes. This is known as the '*classical ball-plate problem*'.

In the classical ball-plate problem, any rotation of the ball around the vertical axis is usually forbidden. In some cases, the ball rotates between two horizontal plates: The upper plate moves in horizontal direction, thus providing the actuation for the ball. Since the ball has no actuation of its own, especially no actuation around the vertical axis, it is practical to set the vertical rotation rate (called often *spinning*) zero. In the related literature, the complete formulae include also the spinning motion, while several examples in the same references assume no-spinning for convenience. (Cai and Roth 1987; Montana 1988; Li and Canny 1990; Jurdjevic 1993; Bicchi et al. 1995). Jurdjevic writes '*It is convenient to assume that $\omega_3 = 0$.*' On this basis, the *no-slip –constraint* is often defined as rolling without slipping, but also setting the spinning rate zero.

Some robot designs are able to rotate in any direction. Therefore, these *omni-directional ball robots* do not require vertical precession for steering. In modelling of these robots, it is beneficial to assume that the contact geometry and friction prevent spinning. (Halme et al. 1996; Bicchi et al. 1997; Spitzmüller 1998; Camicia et al. 2000; Alves and Dias 2003; Chen et al. 2012) For their omni-directional robot and based on no-spin –condition, Zhan et al. (2011) set total vertical rotation rate as zero. From the rolling velocity, lean angle and no-spin constraint they solve the ZYX-Euler rotation velocities. The result particularly describes the no-spin condition: the ball does not rotate around the vertical axis, but it rotates only around a horizontal axis.

Mukherjee et al. (1999, 2002) and Das and Mukherjee (2004, 2006) study the application of the classical ball-plate problem for path planning and steering of a ball robot. In addition, they introduce a model of a ball with a tilted rolling axis. The no-slip constraint is set effective including the no-spinning condition. Das

The following calculations –based on the contact kinematics, solve the path radius, the center of the circular path, and the precession velocity. Note that in Figure 3, the circular path center O is not yet fixed with respect to the ball. The path center location becomes known only after the radiuses r_1 and r_2 are solved.

Under no-slip –conditions, the roller velocity at any contact point is zero. Equation (1) presents the kinematic condition at the extreme points of the contact area.

$$\begin{cases} -\omega R'_1 + \Omega r_1 = 0 \\ -\omega R'_2 + \Omega r_2 = 0 \end{cases} \quad (1)$$

Notifying the roller geometry, (1) is identical with

$$\begin{cases} \omega R_1 \cos \theta = \Omega r_1 \\ \omega \left(\frac{R_1 \cos \theta + d \sin \theta}{r_2} \right) = \Omega \left(\frac{r_1 + d}{r_2} \right). \end{cases} \quad (2)$$

In this kinematically constrained situation, the precession rate Ω can be solved from (2) as a function of the rolling rate ω and the lean angle θ .

$$\Omega = \omega \sin \theta \quad (3)$$

Substituting (3) into (2) gives the velocity-independent kinematic equations (4) for solving the path radius.

$$\begin{cases} \omega R_1 \cos \theta = \omega \sin \theta r_1 \\ \omega R_2 \cos \theta = \omega \sin \theta r_2 \end{cases} \quad (4)$$

Equation (4) can be expressed as

$$\frac{R_1}{r_1} = \frac{R_2}{r_2} = \tan \theta. \quad (5)$$

Equation (5) applies to all points on the contact area. As r_c presents the path radius measured to the center of the contact area, (5) can be written as

$$r_c = \frac{R}{\tan \theta}. \quad (6)$$

Equation (6) indicates that the rolling axis ζ of the cone indeed passes through the path center O .

In case of a finite contact area, this result gives answers to our three first questions: In such case, the no-slip condition and contact kinematics invoke the precession motion and make the ball to follow a circular path, whose center is located at the point where the extension of the tilted rolling axis meets the rolling plane.

The next question is: Can this result be extended to apply also for a ball with a point contact? To answer to this question, we have a look at the calculations we have made: The kinematically constrained precession rate (3) was derived by solving the equation pair in (2). Re-arranging of (2) gives

$$\begin{cases} \omega R_1 \cos \theta = \Omega r_1 \\ \omega R_1 \cos \theta + \omega d \sin \theta = \Omega r_1 + \Omega d \end{cases} \quad (7)$$

$$\Leftrightarrow \begin{cases} \omega R_1 \cos \theta - \Omega r_1 = 0 \\ \omega R_1 \cos \theta - \Omega r_1 = d(\Omega - \omega \sin \theta) \end{cases} \quad (8)$$

If the length d was zero, i.e. the contact comprises a point, (8) is equivalent with (9).

$$\Leftrightarrow \begin{cases} \omega R_1 \cos \theta - \Omega r_1 = 0 \\ 0 = 0 \end{cases}, d = 0 \quad (9)$$

The lower row of (9) is identically true and the two unknowns (path radius and precession rate) can't be solved with the one remaining equation in (9). On basis of this result, the conical roller analogy can't be extended for a point contact. The no-slip –condition still prevails, but it does not define the precession rate of the ball. Path radius of a rolling ball can then be different from what would be expected on based on the kinematic model. Such point contact could be present, for example, at the contact of a rolling disc or a rolling ideal sphere.

ROLLING PATH RADIUS DERIVED FROM BALL DYNAMICS

An ideal sphere with a point contact may be encountered for example in simulation experiments of ball-shaped robots, especially in multi-body simulation environments, like Adams. As was discovered in the previous section, in this particular case the kinematic model can't be used to define the path radius, and some other solution must be found. This section seeks to solve the precession rate from ball dynamics.

Figure 4 presents a ball with a point contact. A zero torque T_z indicates the lack of the vertical friction torque, which is due to the point contact that does not provide any moment arm for lateral friction forces. Path radius r and precession velocity Ω are unknown. The forward rolling axis ζ crosses the rolling plane at point P located at the distance of $R/\tan \theta$ from the contact point. Note that we have not placed the path center O in any specific location with respect to the ball center. Distance from the crossing point P to the path center O depends solely on the path radius r , which is yet unknown. This means also that the crossing point P on the rolling plane is not necessarily stationary.

The general rolling constraint presented in (1) remains valid and it can be expressed as in (10). Path radius r can be solved from (10) once the precession rate Ω is known.

$$\omega R' \equiv \omega R \cos \theta = \Omega r \quad (10)$$

Equations of motion for a rotating body provide the necessary tools for solving the precession rate. Equation (11) presents Euler's equations for a rotationally

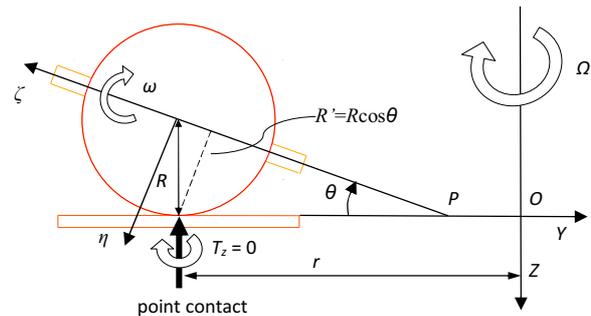


Figure 4: A Rolling Ball with a Point Contact

symmetric object, adapted from Hibbeler (2009, p. 603-615) to apply the notation of Figure 4. This particular form of Euler's equations has been modified for analysis of a spinning top. The moving frame of reference follows the nutation and precession of the top, but does not spin with it. In particular, the accelerations need to be determined in the moving frame.

$$\begin{aligned} M_\xi &= I(\ddot{\theta} + \Omega^2 \sin \theta \cos \theta) + I_0 \Omega \cos \theta (-\Omega \sin \theta + \omega) \\ M_\eta &= I(\dot{\Omega} \cos \theta - 2\Omega \dot{\theta} \sin \theta) - I_0 \dot{\theta} (-\Omega \sin \theta + \omega) \quad (11) \\ M_\zeta &= I_0(\dot{\omega} - \dot{\Omega} \sin \theta - \Omega \dot{\theta} \cos \theta) \end{aligned}$$

The left-hand side of (11) presents the torques affecting the ball, and the right-hand side presents its motion and inertia terms. I_0 stands for the inertia around the rolling axis ζ , and I indicates the inertia around the principal axes η and ξ . The inertias I_0 and I are not assumed to be identical, but they may differ, for example due to the mechanical structure of the robot's spherical shell. Equation (12) presents torque balance around the vertical axis.

$$M_\eta \cos \theta - M_\zeta \sin \theta - T_Z = 0 \quad (12)$$

Substituting (11) into (12) with consideration of ball inertia about vertical axis gives precession acceleration:

$$\dot{\Omega} = \frac{T_Z + 2\Omega \dot{\theta} \sin \theta \cos \theta (I - I_0) + I_0(\dot{\theta} \omega \cos \theta + \dot{\omega} \sin \theta)}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}} \quad (13)$$

Equation (13) presents a dynamic model for the ball precession. The model includes rolling velocity and the lean angle of the robot, their time derivatives, and the two main moments of inertia.

Recall that Equations (11) and (12) represent the spherical shell alone neglecting any other parts of the robot. For completeness, the denominator of (13) has been added with the driving mechanism inertia I_{pz} around the ground vertical. Exact pendulum angle and inertia I_{pz} are functions of ball precession velocity, lean angle and path radius. For brevity, a theoretical scalar-valued maximum inertia of the pendulum presents a worst-case estimate for I_{pz} . In the following, the simulation results are repeated with and without the pendulum inertia. We may also note that in absence of pendulum inertia, the model in (13) is general for rolling spheres and applicable for many different types of ball-robots. Only requirement is that the shell rotates around a tilted rolling axis. This model can then be completed with the specific inertial properties of the robot being under investigation.

For demonstration purposes, the model in (13) can be simplified to present an ideally controlled ball having a constant lean angle θ . In case of a constant lean angle, the precession acceleration becomes

$$\dot{\Omega} = \frac{T_Z + I_0 \dot{\omega} \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}}, \dot{\theta} = 0 \quad (14)$$

The precession acceleration is now a function of the ball rolling acceleration. In absence of external torque and during a steady-state rolling, precession rate remains constant and the ball follows a steady circular path.

Upon ball acceleration, precession rate develops respectively and the ball maintains the same circular path. During acceleration from rest to the final rolling velocity ω , the ball precession velocity Ω can be acquired through integration of (14). In absence of external torque and assuming a constant worst-case I_{pz}

$$\Omega = \frac{I_0 \omega \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}}, T_Z = 0 \quad (15)$$

The path radius can then be solved from (10) and (15)

$$\omega R \cos \theta = \frac{I_0 \omega \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}} \cdot r, \quad (16)$$

$$\Leftrightarrow r = \left(\frac{I}{I_0} \cos^2 \theta + \sin^2 \theta + \frac{I_{pz}}{I_0} \right) \cdot \frac{R}{\tan \theta}. \quad (17)$$

Equation (17) presents the path radius for a ball that maintains a constant lean angle, being independent from ball velocity or acceleration. It is interesting to note that in case of a uniform sphere without any pendulum, $I = I_0$, $I_{pz} = 0$ and (17) is equivalent with (6), and (15) equals to (3). This means that, under no-slip conditions, the dynamic behavior of a uniform ball with a point contact is similar to the kinematic behavior of a ball with a finite contact area. In general and due to differences in main moments of inertia, this is not the case. For a constant lean angle, comparison of the path radius from the dynamic model (17) with the kinematic model (6) reveals the ratio

$$r_{kinematic}/r_{dynamic} = \left(\frac{I}{I_0} \cos^2 \theta + \sin^2 \theta + \frac{I_{pz}}{I_0} \right) \quad (18)$$

In practical applications, the robot's main moment of inertia I_0 around the forward rolling axis is often smaller than the inertia I about the two other axes. As a practical example, the GimBall-robot robot shell in Figure 1 has the inertia ratio 1.18. Assuming a 0.2 rad lean angle and a point contact, the dynamic model attains a path radius 17% larger than suggested by the kinematic model. GimBall pendulum inertia may enlarge the path radius up to 41% beyond the kinematic model.

SIMULATION AND RESULTS

For simulation purposes, the precession acceleration and rate can be calculated in a closed form from (14) and (15). In this example, constant lean angle and rolling acceleration are applied for a given acceleration time T_a

$$\dot{\omega}(t) = \begin{cases} \dot{\omega}, & 0 \leq t \leq T_a \\ 0, & t > T_a \end{cases}. \quad (19)$$

Equation (20) shows the resulting rolling velocity ω , and the precession acceleration comes from (14) according to (21).

$$\omega(t) = \begin{cases} t \dot{\omega}, & 0 \leq t \leq T_a \\ T_a \dot{\omega}, & t > T_a \end{cases} \quad (20)$$

$$\dot{\Omega}(t) = \begin{cases} \frac{I_0 \dot{\omega} \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}}, & 0 \leq t \leq T_a \\ 0, & t > T_a \end{cases} \quad (21)$$

Integration of (21) gives the precession velocity Ω in (22). Further integration gives the ball heading φ

according to (23).

$$\Omega(t) = \int_0^t \dot{\Omega}(t) dt = \int_0^t \frac{I_0 \dot{\omega} \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}} dt =$$

$$= \begin{cases} \frac{I_0 \dot{\omega} t \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}}, & 0 \leq t \leq T_a \\ \frac{I_0 \dot{\omega} T_a \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}}, & t > T_a \end{cases} \quad (22)$$

$$\varphi(t) = \int_0^t \Omega(t) dt = \int_0^t \left(\int_0^t \frac{I_0 \dot{\omega} t \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}} dt \right) dt =$$

$$= \begin{cases} \frac{I_0 \dot{\omega} t^2 \sin \theta}{2(I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz})}, & 0 \leq t \leq T_a \\ \frac{I_0 \dot{\omega} T_a \sin \theta}{I \cos^2 \theta + I_0 \sin^2 \theta + I_{pz}} \cdot \left(t - \frac{T_a}{2} \right), & t > T_a \end{cases} \quad (23)$$

Knowing the rolling velocity ω and heading φ , the ball velocity components v_x and v_y can be calculated as

$$\begin{cases} v_x = R\omega(t)\cos\theta \cdot \cos\varphi(t) \\ v_y = R\omega(t)\cos\theta \cdot \sin\varphi(t) \end{cases} \Leftrightarrow \quad (24)$$

$$v_x = \begin{cases} Rt\dot{\omega}\cos\theta\cos\left(\frac{I_0\dot{\omega}t^2\sin\theta}{2(I\cos^2\theta + I_0\sin^2\theta + I_{pz})}\right), & 0 \leq t \leq T_a \\ RT_a\dot{\omega}\cos\theta\cos\left(\frac{I_0\dot{\omega}T_a\sin\theta \cdot \left(t - \frac{T_a}{2}\right)}{I\cos^2\theta + I_0\sin^2\theta + I_{pz}}\right), & t > T_a \end{cases}$$

$$v_y = \begin{cases} Rt\dot{\omega}\cos\theta\sin\left(\frac{I_0\dot{\omega}t^2\sin\theta}{2(I\cos^2\theta + I_0\sin^2\theta + I_{pz})}\right), & 0 \leq t \leq T_a \\ RT_a\dot{\omega}\cos\theta\sin\left(\frac{I_0\dot{\omega}T_a\sin\theta \cdot \left(t - \frac{T_a}{2}\right)}{I\cos^2\theta + I_0\sin^2\theta + I_{pz}}\right), & t > T_a \end{cases} \quad (25)$$

An adaptive Gauss-Kronrod quadrature function, provided by Matlab software, was used to numerically integrate the ball's location from the velocities in (25), applying the default values for absolute error tolerance (1E-10) and relative error tolerance (1E-6).

Figure 5 demonstrates ball paths with varying inertia ratios, but with a constant 0.2-rad lean angle and constant 1-rad/s² acceleration to reach the 1-rad/s forward rolling velocity at $t = 1$ s. Ball radius is 0.226 m. The precession motion develops during the 1-s acceleration period after which the ball continues rolling with constant velocity and precession rate until the end of simulation at $t = 100$ s. Two separate results are produced with parallel simulations: one in Matlab using the equations derived above and marked with 'o', and another in Adams multi-body simulation software and marked with 'x'. The figure shows no differences in the parallel simulation results. During a 100-s simulation the ball rolled 22 m, while the maximum measured location difference between the two simulation results remains below 0.00013 mm. The difference is less than $6 \cdot 10^{-7}$ percent of the travelled distance.

As well in Figure 5, a third simulation result marked with 'Δ' was added to demonstrate the effect from the theoretical maximum GimBall pendulum inertia around the vertical axis. In this simulation, shell rolling inertia is 0.0633 kg·m², pendulum mass 1.795 kg, pendulum length 0.065 m and max. pendulum main moment of inertia 0.0074 kg·m², as derived from a 3D-model. The added 0.015 kg·m²-pendulum inertia makes the

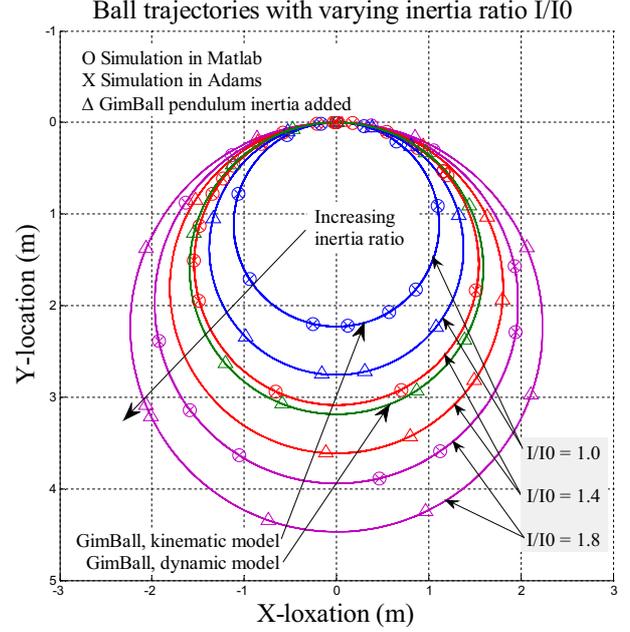


Figure 5: Ball Trajectories with Varying Inertia Ratio, with and without the Added GimBall Pendulum Inertia. Ball Radius 0.224 m, Lean Angle 0.2 rad

simulated rolling paths significantly larger.

The simulation results in Figure 5 demonstrate how different inertia ratios lead to different paths, even though the lean angle, the rolling velocity acceleration, and the final velocity are the same. This is a clear difference to the kinematic precession model. The presence of the pendulum makes the difference even larger. The innermost path in Figure 5 describes the inertia ratio 1, being similar to the result from the kinematic constraint. The difference between this path and the other paths reveals the different behavior of the dynamic and kinematic models.

COMBINING THE KINEMATIC AND DYNAMIC MODELS

We have now discovered that, with a finite contact area, a kinematic precession model can be used to define the radius of the circular path, as long as the no-slip condition prevails. We have also learned that there exists a gyroscopic torque that creates a precession motion even in absence of the kinematic rolling constraint. Now, it is of interest to study the circumstances where the kinematic rolling constraint remains valid.

The kinematic precession model requires that the contact friction provides the necessary forces to maintain (and also to change) the ball state of motion. The friction forces are needed to accelerate the ball velocity, to prevent sliding under centrifugal force, and as well to create the vertical torque to change the precession rate, if not created by the gyroscopic forces as shown in (13). Depending on the inertia ratio, gyroscopic precession torque may increase or decrease the needed external friction torque. In order to create the

sufficient torque, the contact geometry needs to have a sufficient diameter. In this section, based on the kinematic and dynamic precession models, we calculate the minimum necessary contact area diameter. With this contact area, the kinematic precession model remains valid and the ball follows the anticipated circular path.

In case of a finite contact area depicted in Figure 3, the time derivative of the kinematic equation (3) connects the ball precession rate change to the ball forward rolling acceleration

$$\dot{\Omega} = \dot{\omega} \sin \theta. \quad (26)$$

With a given lean angle θ , the dynamic equation (14) presents the precession acceleration as a function of the vertical friction torque and precession torque. The necessary friction torque can then be solved

$$T_z = \dot{\omega} \sin \theta \cdot \left[I_0 \left(\frac{I}{I_0} - 1 \right) \cos^2 \theta + I_{pz} \right]. \quad (27)$$

In (27) one may note that a symmetric ball (without a pendulum) would not need any external friction torque. In such case the gyroscopic torque on the ball creates the necessary precession acceleration. We may see also that the friction torque is needed only when the ball rolling velocity changes.

In case of the GimBall-robot, the needed external precession torque for the shell at 0.2 rad lean angle and 1 rad/s forward rolling angular acceleration is 2.17 mNm. Addition of the maximum theoretical pendulum inertia increases the torque need to 5.15 mNm.

The contact pressure is assumed to be uniform over the circular contact area. The friction force over the surface builds up of two components: a) Lateral acceleration friction force counteracts the forward acceleration and centrifugal acceleration force. This force distributes evenly over the contact area and has the same direction at every point. b) Vertical friction torque develops from lateral force components that have uniform magnitude, but direction vectors rotate around the contact point. To prevent slipping, the maximum combined friction force shall not exceed the contact friction created by the ball weight and the coefficient of friction.

Respectively with the two force components mentioned above, the contact friction coefficient is divided into two components: μ_a covers the lateral acceleration force, and μ_T covers the vertical friction torque. To prevent slipping, the total friction coefficient needed is the sum of these

$$\mu = \mu_a + \mu_T. \quad (28)$$

Friction force providing the lateral acceleration is independent from the contact area. The necessary lateral friction coefficient can be calculated from the lateral ball acceleration and the centrifugal acceleration as

$$(mg\mu_a)^2 = (mR\dot{\omega} \cos \theta)^2 + (mr\Omega^2)^2 \quad (29)$$

Since the motion follows a circular trajectory, the precession rate and path radius can be inserted from (3) and (6):

$$(mg\mu_a)^2 = (mR\dot{\omega} \cos \theta)^2 + \left(m \frac{R}{\tan \theta} (\omega \sin \theta)^2 \right)^2 \quad (30)$$

$$\Leftrightarrow (mg\mu_a)^2 = (mR\dot{\omega} \cos \theta)^2 + (mR \cos \theta \omega^2 \sin \theta)^2 \quad (31)$$

The friction force needed for the given vertical precession torque depends on the contact area. The torque can be calculated as an integral over the contact area A

$$T_z = 4 \int_0^{\frac{\pi}{2}} \int_0^{\frac{d}{2}} (\underbrace{\mu_T}_{\text{contact pressure}} \cdot \frac{4mg}{\pi d^2} \cdot \rho \cdot \underbrace{d\rho \cdot \rho d\alpha}_{dA}), \quad (32)$$

where m = ball mass, ρ = distance from contact area center, and α = angular location of the force element. This gives

$$T_z = \frac{\mu_T mgd}{3} \Leftrightarrow \mu_T = \frac{3T_z}{mgd}. \quad (33)$$

Equation (33) relates the vertical friction torque to the contact area diameter, normal contact force, and friction coefficient. The same model is applied in CONTACT-statement of Adams simulation software. The necessary contact diameter can be solved from (28), (31) and (33):

$$\mu = \frac{R\sqrt{(\dot{\omega} \cos \theta)^2 + (\cos \theta \omega^2 \sin \theta)^2}}{g} + \frac{3T_z}{mgd} \Leftrightarrow d = \frac{3T_z}{m(\mu g - R\sqrt{(\dot{\omega} \cos \theta)^2 + (\cos \theta \omega^2 \sin \theta)^2})} \quad (34)$$

Inserting the necessary friction torque T_z from (27) to (34) gives the needed contact diameter:

$$d = \frac{3\dot{\omega} \sin \theta \cdot [I_0 \left(\frac{I}{I_0} - 1 \right) \cos^2 \theta + I_{pz}]}{m(\mu g - R\sqrt{(\dot{\omega} \cos \theta)^2 + (\cos \theta \omega^2 \sin \theta)^2})} \quad (35)$$

In (35) one may note that in a symmetric case ($I=I_0$) and in absence of the pendulum inertia, $d = 0$, i.e. a symmetric ball with a point contact will satisfy the constraints. This is because the gyroscopic torque on the symmetric ball creates the necessary precession acceleration. No external contact friction torque is needed. In a non-symmetric case a non-zero contact diameter is needed.

Assuming a coefficient of friction $\mu = 0.2$ (typical for acrylic material used for the GimBall shell), ball radius 0.226 m, GimBall mass $m = 5.1$ kg, inertia ratio 1.18, and $I_0 = 0.0633$ kg·m², the needed contact area diameter becomes $d = 0.75$ mm at the given 1-rad/s forward rolling velocity, 1-rad/s² angular acceleration, and 0.2-rad sideways lean angle. With the maximum theoretical pendulum inertia 0.015 kg·m² added, the needed contact area diameter is 1.76 mm.

Figure 6 presents the result from (35) as a function of ball velocity and friction coefficient. As the centrifugal acceleration increases along with the increased rolling velocity, the remaining friction force available for the vertical friction torque decreases. Therefore, the necessary contact area increases. As the needed contact area diameter escapes to infinity, the forward acceleration and centrifugal acceleration consume all

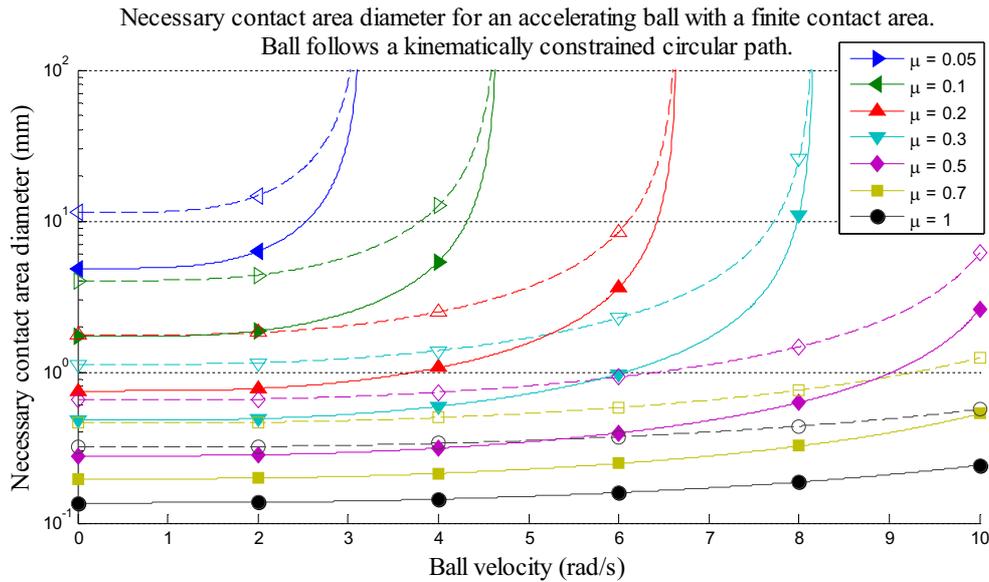


Figure 6: Necessary Contact Area Diameter for an Accelerating GimBall with Different Contact Friction Coefficients. Rolling Acceleration 1 rad/s^2 , Lean Angle 0.2 rad , Path Radius 1.1 m . Dashed lines with open markers indicate the maximum effect of the GimBall pendulum inertia

available friction force capacity, and no friction torque to maintain the precession acceleration remains left. Added inertia from the elevated pendulum increases the necessary contact area. The curved graphs indicate that, due to necessary precession torque, slipping may start earlier than anticipated from rolling acceleration and centrifugal acceleration alone.

The results in Figure 6 assume that the lean angle and path radius remain constant. In addition to contact properties, the ability of the robot to follow this path depends also on the steering mechanism. In case of the GimBall-robot, the short pendulum length limits the maximum rolling velocity along this path to 1.05 m/s . Therefore, GimBall loses the track rather by flipping aside than sliding. Slippage becomes evident only at low coefficients of friction or at higher acceleration.

CONCLUSIONS

This study has discussed the kinematic and dynamic precession models for a rolling ball with a finite contact area and a point contact respectively. In literature, both conventions have been applied.

Usually, literature does not define the contact geometry. Instead, ideal spheres and planes are regularly introduced without mentioning a possibility for deformation or a finite contact area. Reader is then tempted to assume that the presented models in literature comprise a point contact. However, these models often employ the kinematic precession model, which applies for a point contact only in a case of an ideal symmetric sphere where $I=I_0$. In practical applications, such ideal symmetry can rarely be assumed.

In multi-body dynamics simulators, like Adams, the selection of the finite contact area versus a point contact makes a difference. In Adams, the GCON (general

constraint) –statement can be used to create an ideal point contact. For a ball with the point contact, the dynamic precession model then prevails instead of the kinematic one. In contrast, the CONTACT –statement in Adams creates a contact with a finite contact area. In this case, the kinematic precession model applies, -as long as contact friction is sufficient. If the simulator model presents an ideal point contact but a kinematic behavior is desired, it is necessary to implement additional artificial constraints to invoke the kinematic precession model.

We have introduced the dynamic precession model for a rolling ball-robot. This model explains ball precession in presence of an ideal point contact. The absence of pendulum parameters allows application of the model for different types of ball-shaped robots. The model can be used in simulators, and it provides also information about the gyroscopic precession torque.

We have applied both the kinematic and dynamic precession model to evaluate the no-slip condition of the GimBall-robot. The result indicates that ball slipping may start earlier than anticipated if neglecting the need for an external precession torque.

In the progressive studies, the scalar-valued worst-case pendulum inertia estimate I_{pz} may be replaced with an exact presentation taking into consideration the actual pendulum orientation. Experimental tests call for robots with especially rigid spherical shells to demonstrate the effect of a minute contact area on a slippery surface.

REFERENCES

- Alves, J. and Dias, J. 2003. "Design and control of a spherical mobile robot." *Proceedings Of The Institution Of Mechanical Engineers Part I-Journal of Systems and Control Engineering* 217, No.6, 457-467.

- Avery, G. 2011. "Sphero rolling out as orders tumble in." *Denver Business Journal*, Dec 2, 2011, A1.
- Azizi, M.R. and Naderi, D. 2013. "Dynamic modeling and trajectory planning for a mobile spherical robot with a 3Dof inner mechanism." *Mechanism and Machine Theory* 64, 251-261.
- Bhattacharya, S. and Agrawal, S.K. 2000a. "Design, experiments and motion planning of a spherical rolling robot." In *Proceedings of IEEE International Conference on Robotics and Automation* (San Francisco, CA, April 2000). IEEE, 1207-1212.
- Bhattacharya, S. and Agrawal, S.K. 2000b. "Spherical rolling robot: a design and motion planning studies." *IEEE Transactions on Robotics and Automation* 16, No.6, 835-839.
- Bicchi, A.; Balluchi, A.; Prattichizzo, D. and Gorelli, A. 1997. "Introducing the "SPHERICLE": an experimental testbed for research and teaching in nonholonomy." In *Proceedings of IEEE International Conference on Robotics and Automation* (Albuquerque, NM, April 1997). 2620-2625.
- Bicchi, A.; Prattichizzo, D. and Sastry, S.S. 1995. "Planning motions of rolling surfaces." In *Proceedings of IEEE Conference on Decision and Control* (New Orleans, Dec. 1995). 2812-2817.
- Brown Jr., H.B. and Xu, Y. 1996. "A single-wheel, gyroscopically stabilized robot." In *Proceedings of IEEE International Conference on Robotics and Automation* (Minneapolis, MN, April 1996). IEEE, 3658-3663.
- Bruhn, F.C.; Pauly, K. and Kaznov, V. 2005. "Extremely low mass spherical rovers for extreme environments and planetary exploration enabled with MEMS." *European Space Agency, (Special Publication) ESA SP 603*, 347-354.
- Cai, C. and Roth, B. 1987. "On the spatial motion of a rigid body with point contact." In *Proceedings of IEEE International Conference on Robotics and Automation* (Raleigh, NC, Mar. 1987). IEEE, 686-695.
- Cai, Y.; Zhan, Q. and Xi, X. 2012. "Path tracking control of a spherical mobile robot." *Mechanism and Machine Theory* 51, 58-73.
- Camicia, C.; Conticelli, F. and Bicchi, A. 2000. "Nonholonomic kinematics and dynamics of the Sphericle." In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (Takamatsu, Japan, Nov. 2000). 805-810.
- Chen, W.; Chen, C.; Yu, W.; Lin, C. and Lin, P. 2012. "Design and implementation of an omnidirectional spherical robot Omnicron." In *Proceedings of IEEE/ASME International Conference on Advanced Intelligent Mechatronics* (Kachsiung, July 2012). IEEE, 719-724.
- Das, T. and Mukherjee, R. 2004. "Exponential stabilization of the rolling sphere." *Automatica* 40, No.11, 1877-1889.
- Das, T. and Mukherjee, R. 2006. "Reconfiguration of a rolling sphere: A problem in evolute-involute geometry." *ASME Journal of Applied Mechanics* 73, No.4, 590-597.
- Ghanbari, A.; Mahboubi, S. and Fakhrabadi, M.M.S. 2010. "Design, dynamic modeling and simulation of a spherical mobile robot with a novel motion mechanism." In *Proceedings of IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications* (Qingdao, Shan Dong, July 2010). 434-439.
- Halme, A.; Schonberg, T. and Wang, Y. 1996. "Motion control of a spherical mobile robot." In *Proceedings of IEEE International Workshop on Advanced Motion Control* (Mie, 18-21 Mar.). 259-264.
- Hibbeler, R.C. 2009. *Engineering Mechanics: Combined Statics & Dynamics*. 12th ed. Prentice Hall, Upper Saddle River, NJ.
- Hristu-Varsakelis, D. 2001. "The dynamics of a forced sphere-plate mechanical system." *IEEE Transactions on Automatic Control* 46, No.5, 678-686.
- Ishikawa, M.; Kitayoshi, R. and Sugie, T. 2010. "Dynamic rolling locomotion by spherical mobile robots considering its generalized momentum." In *Proceedings of Society of Instrument and Control Engineers Annual Conference* (Taipei, Aug. 18-21). IEEE, 2311-2316.
- Ishikawa, M.; Kitayoshi, R. and Sugie, T. 2011. "Volvo: A spherical mobile robot with eccentric twin rotors." In *Proceedings of IEEE International Conference on Robotics and Biomimetics* (Karon Beach, Phuket, Dec. 2011). IEEE, 1462-1467.
- Javadi, A.H.A. and Mojabi, P. 2002. "Introducing August: a novel strategy for an omnidirectional spherical rolling robot." In *Proceedings of IEEE International Conference on Robotics and Automation* (Washington, DC, May 2002). IEEE, 3527-3533.
- Javadi, A.H.A. and Mojabi, P. 2004. "Introducing Glory: A novel strategy for an omnidirectional spherical rolling robot." *Journal of Dynamic Systems, Measurement and Control, Transactions of the ASME* 126, No.3, 678-683.
- Jia, Q.; Sun, H. and Liu, D. 2008. "Analysis of Actuation for a Spherical Robot." In *Proceedings of IEEE Conference on Robotics, Automation and Mechatronics* (Chengdu, China, 21-24 Sept.). 266-271.
- Joshi, V.A. and Banavar, R.N. 2009. "Motion analysis of a spherical mobile robot." *Robotica* 27, No.3, 343-353.
- Joshi, V.A.; Banavar, R.N. and Hippalgaonkar, R. 2007. "Design, modeling and controllability of a spherical mobile robot." In *Proceedings of National Conference on Mechanisms and Machines* (Bangalore, India, Dec. 2007). 135-140.
- Joshi, V.A.; Banavar, R.N. and Hippalgaonkar, R. 2010. "Design and analysis of a spherical mobile robot." *Mechanism and Machine Theory* 45, No.2, 130-136.
- Jurdjevic, V. 1993. "The geometry of the plate-ball problem." *Archive for Rational Mechanics and Analysis* 124, No.4, 305-328.
- Kamaladar, M.; Mahjoob, M.J.; Haeri Yazdi, M.; Vahid-Alizadeh, H. and Ahmadzadeh, S. 2011. "A control synthesis for reducing lateral oscillations of a spherical robot." In *Proceedings of IEEE International Conference on Mechatronics* (Istanbul, Apr. 2011). 546-551.
- Kayacan, E.; Bayraktaroglu, Z.Y. and Saeys, W. 2012. "Modeling and control of a spherical rolling robot: a decoupled dynamics approach." *Robotica* 30, No.4, 671-680.
- Kaznov, V. and Seeman, M. 2010. "Outdoor navigation with a spherical amphibious robot." In *Proceedings of Intelligent Robots and Systems* (Taipei, Oct. 2010). 5113-5118.
- Kim, J.; Kwon, H. and Lee, J. 2009. "A rolling robot: Design and implementation." In *Proceedings of 7th Asian Control Conference* (Hong Kong, Aug. 2009). 1474-1479.
- Krieger, K. 2013. "Meet GuardBot!." *Physics Today Online*, Aug. 2013, <http://dx.doi.org/10.1063/PT.4.2536>
- Laplanche, J.; Masson, P. and Michaud, F., 2007. *Analytical Longitudinal and Lateral Models of a Spherical Rolling Robot*. [online]. Available at: <http://introlab.3it.usherbrooke.ca/papers/TRRoball.pdf> (Accessed 10.5.2013).
- Li, Z. and Canny, J. 1990. "Motion of two rigid bodies with rolling constraint." *IEEE Transactions on Robotics and Automation* 6, No.1, 62-72.

- Liu, D. and Sun, H. 2010. "Nonlinear sliding-mode control for motion of a spherical robot." In *Proceedings of 29th Chinese Control Conference* (Beijing, 29–31 July). 3244-3249.
- Liu, D.; Sun, H. and Jia, Q. 2008. "Stabilization and path following of a spherical robot." In *Proceedings of IEEE International Conference on Robotics, Automation and Mechatronics* (Chengdu, Sept. 2008). 676-682.
- Mahboubi, S.; Seyyed Fakhrebadi, M.M. and Ghanbari, A. 2013. "Design and implementation of a novel spherical mobile robot." *Journal of Intelligent and Robotic Systems: Theory and Applications* 71, No.1, 43-64.
- Michaud, F. and Caron, S. 2002. "Roball, the rolling robot." *Autonomous Robots* 12, No.2, 211-222.
- Montana, D.J. 1988. "Kinematics of contact and grasp." *International Journal of Robotics Research* 7, No.3, 17-32.
- Morinaga, A.; Svinin, M. and Yamamoto, M. 2012. "On the motion planning problem for a spherical rolling robot driven by two rotors." In *Proceedings of IEEE/SICE International Symposium on System Integration* (Fukuoka, Dec. 2012). IEEE, 704-709.
- Mukherjee, R.; Minor, M.A. and Pukrushpan, J.T. 1999. "Simple motion planning strategies for spherobot: a spherical mobile robot." In *Proceedings of the 38th IEEE Conference on Decision and Control* (AZ, 1999). 2132-2137.
- Mukherjee, R.; Minor, M.A. and Pukrushpan, J.T. 2002. "Motion planning for a spherical mobile robot: Revisiting the classical ball-plate problem." *ASME Journal of Dynamic Systems, Measurement, and Control* 124, No.4, 502-511.
- Nagai, M. 2008. *Control System of a Ball-shaped Robot*. Master's Thesis, Department of Automation and Systems Technology, Helsinki University of Technology.
- Nandy, G.C. and Xu, Y. 1998. "Dynamic model of a gyroscopic wheel." In *Proceedings of IEEE International Conference on Robotics and Automation* (Leuven, May 1998). IEEE, 2683-2688.
- Otani, T.; Urakubo, T.; Maekawa, S.; Tamaki, H. and Tada, Y. 2006. "Position and attitude control of a spherical rolling robot equipped with a gyro." In *Proceedings of IEEE International Workshop on Advanced Motion Control* (Istanbul, March 27-29). IEEE, 416-421.
- Sang, S.; Shen, D.; Zhao, J.; Xia, W. and An, Q. 2011. "Prototype Design and Motion Analysis of a Spherical Robot." *Communications in Computer and Information Science* 134, No.1, 284-289.
- Sang, S.; Zhao, J.; Wu, H.; Chen, S. and An, Q. 2010. "Modeling and simulation of a spherical mobile robot." *Computer Science and Information Systems* 7, No.1, 51-62.
- Sang, S.J.; Shen, D.; Zhao, J.C.; Hu, J.Y. and An, Q. 2011. "Analysis and Simulation of a Spherical Robot." *Advanced Materials Research* 171 - 172, 748-751.
- Spitzmüller, S. 1998. *Microcontroller based control system for a rolling minirobot*. Master's Thesis, Department of Automation and Systems Technology, Helsinki University of Technology.
- Svinin, M. and Hosoe, S. 2006. "Simple motion planning algorithms for ball-plate systems with limited contact area." In *Proceedings of IEEE International Conference on Robotics and Automation* (Orlando, FL, May 2006). IEEE, 1755-1761.
- Svinin, M. and Hosoe, S. 2008. "Motion Planning Algorithms for a Rolling Sphere With Limited Contact Area." *IEEE Transactions on Robotics* 24, No.3, 612-625.
- Svinin, M.; Morinaga, A. and Yamamoto, M. 2012a. "An analysis of the motion planning problem for a spherical rolling robot driven by internal rotors." In *Proceedings of IEEE International Conference on Intelligent Robots and Systems* (Vilamoura, Oct. 2012). IEEE, 414-419.
- Svinin, M.; Morinaga, A. and Yamamoto, M. 2012b. "On the dynamic model and motion planning for a class of spherical rolling robots." In *Proceedings of IEEE International Conference on Robotics and Automation* (Saint Paul, May 2012). 3226-3231.
- Xu, Y.; Au, K.W.; Nandy, G.C. and Brown, H.B. 1998. "Analysis of actuation and dynamic balancing for a single-wheel robot." In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (Victoria, BC, Oct. 1998). IEEE, 1789-1794.
- Ylikorpi, T. and Suomela, J. 2007. "Ball-shaped Robots." In *Climbing & Walking Robots, Towards New Applications*, Z. Houxiang Zhang (Ed.). InTech, Vienna, 235-256.
- Zhan, Q.; Cai, Y. and Yan, C. 2011. "Design, analysis and experiments of an omni-directional spherical robot." In *Proceedings of IEEE International Conference on Robotics and Automation* (Shanghai, May 2011). IEEE, 4921-4926.
- Zheng, M.; Zhan, Q.; Liu, J. and Cai, Y. 2011. "Control of a Spherical Robot: Path Following Based on Nonholonomic Kinematics and Dynamics." *Chinese Journal of Aeronautics* 24, No.3, 337-345.

AUTHOR BIOGRAPHIES

TOMI YLIKORPI went to Helsinki University of Technology and obtained his Master's degree in 1994 in Mechanical Engineering. He obtained his Licentiate's degree in Automation Technology in 2008 and currently works at Aalto University with design and modelling of mechanical systems. His e-mail address is: TOMI.YLIKORPI@AALTO.FI

PEKKA FORSMAN received his PhD in Automation technology from Helsinki University of Technology in 2001. He is currently university lecturer at Aalto University. His research interests include field and service robotics, human-robot interfaces as well as localization and navigation methods. His e-mail address is: PEKKA.FORSMAN@AALTO.FI

AARNE HALME started his career in 1966 at Helsinki University of Technology. From 1985 until lately he has been the Professor and Head of the Automation Technology Laboratory at Helsinki University of Technology. He is one of the very first pioneers in development and analysis of ball-shaped rolling robots. His e-mail address is: AARNE.HALME@AALTO.FI

UNIFIED REPRESENTATION OF DECOUPLED DYNAMIC MODELS FOR PENDULUM-DRIVEN BALL-SHAPED ROBOTS

Tomi Ylikorpi
Pekka Forsman
Arne Halme
Department of Electrical
Engineering and Automation
Aalto University
P.O. Box 15500
00076 Aalto, Finland
E-mail: tomi.ylikorpi@aalto.fi

Jari Saarinen
GIM – Finnish Centre of Excellence
in Generic Intelligent Machines Research
E-mail: jari.p.saarinen@gmail.com

KEYWORDS

Modelling and simulation in robotic applications, ball-shaped robots, robot dynamics, Euler-Lagrange equation, multi-body simulation

ABSTRACT

Dynamic models describing the ball-robot motion form the basis for developments in ball-robot mechanics and motion control systems. For this paper, we have conducted a literature review of decoupled forward-motion models for pendulum-driven ball-shaped robots. The existing models in the literature apply several different conventions in system definition and parameter notation. Even if describing the same mechanical system, the diversity in conventions leads into dynamic models with different forms. As a result, it is difficult to compare, reproduce and apply the models available in the literature. Based on the literature review, we reformulate all common variations of decoupled dynamic forward-motion models using a unified notation and formulation. We have verified all reformulated models through simulations, and present the simulation results for a selected model. In addition, we demonstrate the different system behavior resulting from different ways to apply the pendulum reaction torque, a variation that can be found in the literature. For anyone working with the ball-robots, the unified compilation of the reformulated dynamic models provides an easy access to the models, as well as to the related work.

INTRODUCTION

Ball-shaped vehicles have been under development already over the last 120 years. The first patents on self-propelled spherical toys were filed in the end of 19th century. Studies on dynamic modelling and steering of motor-driven balls started in 1990's leading into emergence of computer controlled spherical mobile robots. (Ylikorpi and Suomela 2007) Recent studies on ball-shaped robots have described a variety of applications in different environments, including

marine, indoors, outdoors and planetary exploration. Lately, commercial spherical robots have been introduced to the markets. The practical applications include surveillance, rehabilitation and gaming.

Ball-shaped robots offer interesting and challenging modelling and control problems due to their extraordinary dynamic nature. In development of robot mechanics and control, simulation tools play an utterly important role. Simulators regularly represent the robotic system and its behavior, and they are used to verify the performance of the control system. The core of the simulator is the dynamic model describing the ball-robot motion, which also forms the base for the control system development. Thus, the properly formulated dynamic model is of a great importance for development of simulators and control algorithms.

We have conducted a literature review of decoupled forward-motion models for pendulum-driven ball-shaped robots. The survey covered 12 different robots and their models presented in 22 published papers. For describing the robot dynamic model, these publications present several different conventions in system parameters definition and notation, including various model simplifications. This divergence makes it difficult to compare, reproduce and apply the models available in the literature. In this paper, we reformulate in a unified notation all commonly found decoupled forward-motion models of pendulum-driven ball-shaped robots. Our reformulated models, without any simplifications, provide a detailed description of the used assumptions as well as the selected coordinate systems. The unified compilation of the reformulated dynamic models provides an easy access to the existing models.

As is the common practice in the literature, we have verified the performance of each dynamic model through simulations in Matlab-software of MathWorks Inc. (Version 7.5.0.342, R2007b). Additionally, a comparative simulation was performed for each model in Adams multi-body simulation software of MSC.Software Corporation (Version MD Adams R3, Build 2008.1.0). In this context, such model validation

with two parallel and independent simulation tools has been rarely presented before. In addition, we demonstrate the different system behavior resulting from different ways to apply the pendulum reaction torque, a variation that can be found also in the literature. For anyone working with the ball-shaped robots, we present in Appendix 1 the models in a hand-book style providing a clear and easy access to the models, as well as to the related work behind them.

RELATED WORK

Ball-shaped robots represent a family of mobile robots that can be realized with several different mechanisms for actuation, some of which were briefly reviewed by Ylikorpi and Suomela (2007). Plenty of prior work has been conducted on kinematic and dynamic modelling of these robots. Li and Canny (1990), Jurdjovic (1993), and Bicchi et al. (1995) discuss the classical ball-plate – problem. Halme et al. (1996) introduce a ball-robot equipped with a single driving wheel inside the hollow sphere. Bicchi et al. (1997) present another ball robot with a unicycle driving inside the sphere. Camicia et al. (2000) continue the work developing a more advanced dynamic model. Zhan et al. (2011) present another ball-robot based on a unicycle. Mukherjee et al. (1999, 2002) discuss the application of the ball-plate problem, path planning, and steering of a ball robot, while Das and Mukherjee (2004, 2006) develop more complex rolling paths.

Svinin and Hosoe (2008), Svinin et al. (2012a, 2012b), and Morinaga et al. (2012) discuss kinematics, dynamics and control of a ball-robot carrying six flywheels. Karimpour et al. (2012), Joshi et al. (2007, 2010), and Joshi and Banavar (2009) conduct an extensive discussion on a spherical robot driven by three and four momentum wheels.

A motor-actuated hanging pendulum creates one possible driving mechanism, applied for several different ball robots (Koshiyama and Yamafuji 1992, 1993; Michaud and Caron 2002; Bruhn et al. 2005; Kaznov and Seeman 2010; Yoon et al. 2011). Jia et al. (2009), Sang et al. (2011), and Zheng Y.L. (2011) add a momentum wheel on the pendulum.

There are two popular methods to present the equations of motion of a pendulum-driven robot; A coupled model presents the full motion of the complete system. Various mathematical methods, such as Kane’s method, Euler-Lagrange equation, and Maggi’s equations are often applied to create the coupled model (Jia et al. 2008, 2009; Liu et al. 2008; Zhuang et al. 2008; Liu and Sun 2010; Sang et al. 2011; Yu et al. 2011; Zheng, M. et al. 2011; Zheng, Y.L. 2011; Gajbhiye and Banavar 2012; Balandin et al. 2013).

Different from the coupled model, a decoupled model discusses steering and forward-driving motions separately. To mention some methods, decoupled models have been created with application of Newton-Euler-equations, Euler-Lagrange equation, and

Boltzmann-Hamel-equations. We have chosen to apply the Euler-Lagrange equation. Along with our new reformulated models, Appendix 1 presents the reference information for the original works. This survey concentrated on those 22 published models presenting the forward motion of pendulum-driven ball-robots.

COMMON VARIATIONS IN MODEL PRESENTATION

The decoupled forward-motion state of a pendulum-driven ball-robot can be conveniently presented with the ball rotation angle, the pendulum rotation angle, and their time derivatives. The two rotation angles are commonly nominated as the *generalized coordinates* chosen to describe the system state. Ball position along the rolling plane couples directly to the ball rotation through a kinematic rolling constraint.

Figure 1 shows one definition of the generalized coordinates θ_1 and θ_2 , also used by Kim et al. (2009). In this convention, the ball rotation angle θ_1 measures from the ground vertical, and the pendulum elevation angle θ_2 is measured with respect to a reference fixed on the ball. Alternatively, the pendulum angle can be chosen to be the absolute one, while presenting the ball rotation with respect to the pendulum (Koshiyama and Yamafuji 1993). Yet, as one more alternative, both rotation angles can be presented as absolute with respect to the ground (Cai et al. 2011). In addition and opposite to the case shown in Figure 1, the positive rotation direction of the ball can be selected to be clockwise (Yue et al. 2006; Kamaldar et al. 2011; Kayacan et al. 2012a). Table 1 shows the possible permutations available for definition of the two generalized coordinates. All six variations can be found in the literature.

In addition to the different definitions of the generalized coordinates, also the presentation of the ball inertia has different forms. In most of the cases the ball inertia is calculated around the ball center, but some models present the inertia around the contact point (Koshiyama

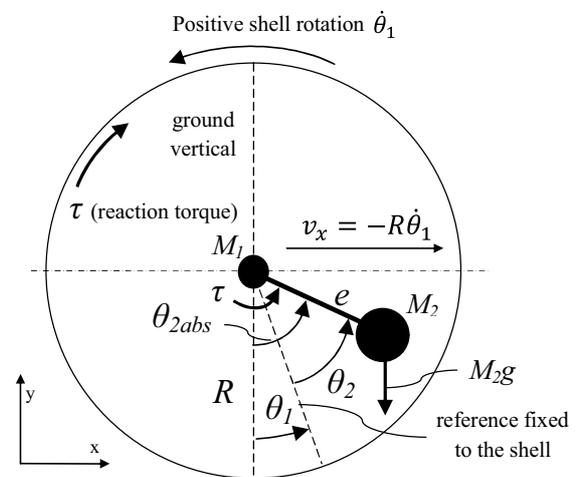


Figure 1: Typical De-coupled Model of a Pendulum-Driven Ball-Robot Moving Forward, Side-View

Table 1: Variants in Definition of the Generalized Angular Coordinates

	Options						Number of options
	Absolute Ball angle, relative Pendulum angle		Absolute Pendulum angle, relative Ball angle		Both absolute		
Selected absolute coordinate							3
Positive rotation direction of the ball with respect to the pendulum	Same	Opposite	Same	Opposite	Same	Opposite	$2 \times 3 = 6$
Illustration in Appendix 1	C)	D)	A)	F)	E)	B)	
Occurrences in the selected literature	2	3	2	1	1	13	

and Yamafuji 1993; Cai et al. 2011; Kamaldar et al. 2011). Our representation calculates the inertia around the ball center.

Different from the other references, Koshiyama and Yamafuji (1992, 1993) present the absolute pendulum rotation angle with respect to the ground horizontal. Our unified representation in Appendix 1 measures the absolute rotation angle from the ground vertical for all models.

The literature presents cases where the pendulum absolute angular velocity is assumed small and the products of the angular velocities can then be neglected (Kim et al. 2009). In addition, sometimes small angles have been assumed thus changing the appearance of trigonometric functions (Kim et al. 2009; Liu et al. 2009). In contrast, we present the complete dynamic equations without simplifications.

Finally, the dynamic model is often presented in a matrix form. Literature shows a couple of different arrangements where the matrix elements and the coordinate vectors are shown in a different order (Koshiyama and Yamafuji 1993; Yue et al. 2006). Appendix 1 presents all models in a uniform arrangement.

THE UNIFIED MODEL

Selection of the Generalized Coordinates

Figure 1 illustrates one possible selection of the generalized coordinates. Ball rotation angle θ_1 is measured counterclockwise from the ground vertical to a reference fixed on the ball. Pendulum rotation angle θ_2 is measured counterclockwise from the reference on the ball towards the pendulum arm. Table 2 explains the parameters and variables used in Figure 1 and in our unified notation.

In the convention shown in Figure 1, the ball angle θ_1 is expressed as an *absolute coordinate*. The absolute coordinate presents directly the ball rotation angle with respect to the inertial world-coordinate system. In contrast, the pendulum rotation angle θ_2 is expressed as a *relative coordinate*. The relative coordinate tells only the pendulum position with respect to the ball. As an alternative presentation for the pendulum orientation,

Figure 1 presents also the absolute pendulum angle θ_{2abs} that measures the pendulum position directly from the ground vertical towards the pendulum arm.

Derivation of the Equations of Motion

The Euler-Lagrange equation can be used to create the equations of motion (Symon 1960; Goldstein et al. 2002), and it has been often applied also with ball-shaped robots (Liu et al. 2008; Jia et al. 2009; Zhang et al. 2009; Kayacan et al. 2012a). Lagrangian function L is defined as the difference between the kinetic energy T and the potential energy V , as shown in (1). Generalized forces Q_i affecting the system can then be solved through derivation of the Lagrangian with respect to time and the generalized coordinates q_i , as presented in (2), known as the Euler-Lagrange equation.

$$L = T - V \tag{1}$$

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = Q_i \tag{2}$$

Referring to the convention in Figure 1, eqs. (3) and (4) present the kinetic and potential energy of the spherical shell and the pendulum. The Lagrangian L in (1) and the differentials on the left side of (2) can then be solved.

Table 2: Parameters for the Dynamic Models in Figure 1 and Appendix 1

M_1 ball mass	M_2 pendulum mass
R ball radius	e pendulum length
J_1 ball inertia	J_2 pendulum inertia
θ_1 ball angle	θ_2 pendulum angle θ_{2abs} absolute pendulum angle
$\dot{\theta}_1$ ball angular velocity	$\dot{\theta}_2$ pendulum angular velocity
c_1 ball rolling friction coefficient	c_2 pendulum joint friction coefficient
T_1 ball kinetic energy	T_2 pendulum kinetic energy
V_1 ball potential energy	V_2 pendulum potential energy
v_x horizontal ball velocity	τ pendulum motor torque

$$T_1 = \frac{1}{2}M_1R^2\dot{\theta}_1^2 + \frac{1}{2}J_1\dot{\theta}_1^2$$

$$T_2 = \frac{1}{2}M_2\left((-R\dot{\theta}_1 + e(\dot{\theta}_1 + \dot{\theta}_2)\cos(\theta_1 + \theta_2))^2\right) \quad (3)$$

$$+ \frac{1}{2}M_2(e(\dot{\theta}_1 + \dot{\theta}_2)\sin(\theta_1 + \theta_2))^2 + \frac{1}{2}J_2(\dot{\theta}_1 + \dot{\theta}_2)^2$$

$$V_1 = 0$$

$$V_2 = -M_2g\cos(\theta_1 + \theta_2) \quad (4)$$

As presented by Koshiyama and Yamafuji (1993), the result from (2) can be expressed in a *configuration space* according to (5), where \mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} and \mathbf{G} denote the matrices including mass and inertia terms, centrifugal terms, coriolis terms, viscous friction, and gravitational forces respectively. Torque vector \mathbf{Q} includes the generalized forces, i.e. the torques affecting the system.

$$\begin{bmatrix} A11 & A12 \\ A21 & A22 \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} B11 & B12 \\ B21 & B22 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1^2 \\ \dot{\theta}_2^2 \end{bmatrix} +$$

$$\begin{bmatrix} C11 \\ C21 \end{bmatrix} (\dot{\theta}_1\dot{\theta}_2) + \begin{bmatrix} D11 & D12 \\ D21 & D22 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} G11 \\ G21 \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \quad (5)$$

The left side of (5) can be acquired through the derivations shown in (2). However, the contents of the torque vector \mathbf{Q} on the right side deserve some discussion, which has not been conducted in the related literature before: The two generalized torques Q_1 and Q_2 relate to the two angular coordinates θ_1 and θ_2 shown in Figure 1. The ball-robot carries a motor that drives the pendulum with respect to shell. The choice of the generalized coordinates defines how the motor torque and its reaction torque project to the generalized torques. Symon (1960, p. 354) notes: ‘...the mutual forces which the particles exert on each other ordinarily depend on the relative coordinate.’ Upon application of the Euler-Lagrange equation, the relative motion between the bodies takes into consideration also the reaction forces between the bodies. Contradictorily, the use of absolute coordinates neglects the reaction forces that then need to be separately taken into account.

To give an example, Figure 1 presents the pendulum angle θ_2 relative to the ball angle. Because of the relative expression of the pendulum angle, the reaction torque from the pendulum motor becomes automatically into consideration through the Euler-Lagrange equation. The pendulum driving torque τ is then included in the system input Q_2 in (5), but the reaction torque is not added explicitly in the ball torque Q_1 . However, if the absolute pendulum angle θ_{2abs} was used instead, the reaction torque must be added also as an input on the ball torque Q_1 .

A proof for the above made statement can be found by calculating symbolically eqs. (1) - (5) using both absolute and relative pendulum angles and notifying the appearance of torques Q_1 and Q_2 in the result. The straightforward calculation is omitted here.

The literature presents both approaches, applying either relative or absolute pendulum coordinate. However, the convention in application of the reaction torque varies. We apply the reaction torque consistently upon need, as is described above, confirmed by the parallel simulations in Adams, and reported in Appendix 1. For comparison, our simulation results present also the different system behavior resulting from the different application of the reaction torque.

Modelling the Viscous Friction

We supplement all dynamic models with viscous friction, which has been previously presented for some formulations. For the given velocity vector $\dot{\boldsymbol{\theta}}$, the manually calculated friction matrix \mathbf{D} provides proper resistance torque for the ball and the pendulum. In the definition of the friction matrix, it is important to note that the frame of reference must be similar to that used in derivation of the Lagrangian in (1). A similar friction model was presented by Koshiyama and Yamafuji (1993), as shown in Appendix 1 A).

Numerical Simulation

All six dynamic models, created with application of (1) and (2), were implemented in Matlab for verification. In the simulation, the dynamic model formulated in configuration space (5) was applied to solve the accelerations for the given input torque \mathbf{Q} . The joint velocities and angles were then integrated with ode45 – solver function. In addition, a parallel model of the system was built in Adams multi-body simulation software and the results were compared for validation.

The ball-robot model is defined for the Adams-software by describing the mechanical structure and the physical properties of the robot. Adams then autonomously creates the dynamic model needed for simulation. Thus, Adams provides a model that is independent from the one created for Matlab, and can be used as a reference in validation of the derived models.

SIMULATION RESULTS

All models collected in Appendix 1 were simulated both in Matlab and in Adams. Comparison of the simulation results verified the correctness of the models. For the sake of brevity, we present the simulation results only for the model according to Figure 1 and applying the formulation C) in Appendix 1.

Figure 2 A) shows the open-loop response for a given pendulum torque impulse. The input torque has a form of a cosine function with a 5-s period and 1-Nm peak value. The integration result in Matlab agrees well with the simulation result in Adams. No difference is visible between the two models in Figure 2 A). Regarding the earlier discussion on the observed variation in application of the reaction torque, the third simulation result in Figure 2 A) reveals the effect from the excess reaction torque in Q_1 . Figure 2 B) demonstrates the

identical behavior of the two independent simulation models; one in Matlab, another in Adams.

Figure 3 repeats the simulations in a closed-loop with a PI-controlled ball velocity. The target ball velocity is -5 rad/s and the controller gains are $P = 0.1$ and $I = 0.3$. In the third simulation, the effect from the excessive reaction torque is clear leading into different conclusion about system dynamics and highly different prediction of the needed pendulum motor torque. The result underlines the importance of the correct dynamic model, being the subject of this paper. Further development and discussion on the control algorithms remain as future work.

Simulations of all model formulations in Appendix 1 produce the same results. In the simulations, the robot model represents the *GimBall*-robot developed at Aalto University having the properties: $M_1 = 3.294$ kg, $M_2 = 1.795$ kg, $R = 0.226$ m, $e = 0.065$ m, $J_1 = 0.0633$ kgm², $J_2 = 0.0074$ kgm², $c_1 = 0.02$ Nms/rad, $c_2 = 0.2$ Nms/rad and $g = 9.81$ m/s².

The models were integrated in Matlab using ode45-solver with the following settings: RelTol = 10^{-6} , AbsTol = 10^{-10} , MaxStep = 10^{-3} and InitialStep = 10^{-6} . The simulator settings in Adams were the corresponding.

CONCLUSIONS AND FURTHER WORK

The dynamic models describing ball-robot motion form the basis for the developments in ball-robot mechanics and motion control algorithms. Thus, the dynamic

model holds an extremely important position in the research on the ball-shaped robots.

Because of the existing diversity in notation and model contents, it is difficult to compare, reproduce and apply the models available in the literature. To facilitate model comparison and re-use, we have in this paper reformulated all common decoupled forward-motion models of pendulum-driven ball-shaped robots. The reformulated models, created with the application of Euler-Lagrange equation while applying a unified notation and harmonized formulation, are collected in Appendix 1.

All reformulated models have been validated with parallel simulations in Matlab and Adams multi-body simulation software. The two independent simulation results show an excellent agreement thus validating the models. Additional simulation results demonstrate the effect from the different conventions in application of reaction torque.

Clear, correct and harmonized description of the dynamic models in a hand-book style is useful for their application in further developments. These models, being the topic of this paper, set the basis for further work of the control algorithms. Our work continues with extension of the dynamic model to consider also other dynamic cases than the pure rolling along a level surface, as well as to consider a robot structure different from an ideal rigid sphere. Practical experiments and addition of the Coulomb friction model are foreseen.

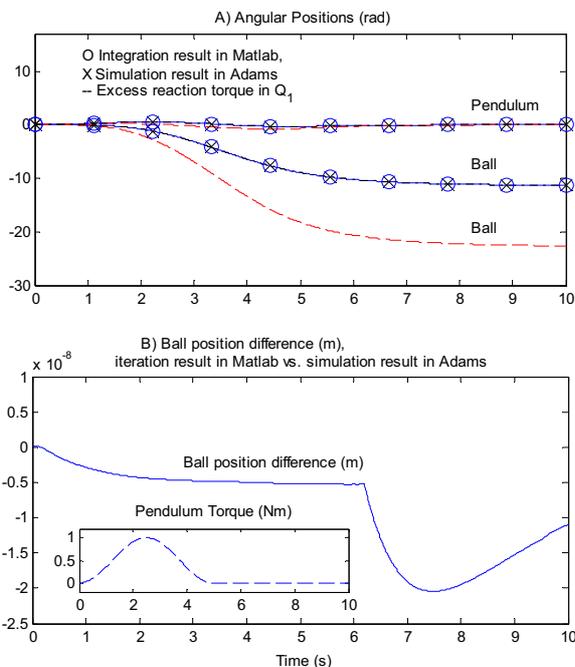


Figure 2: A) In an open-loop simulation, an excess reaction torque in Q_1 causes erroneous system response. B) Ball location difference between the simulation results is $< 2 \cdot 10^{-8}$ m. Inset: the applied pendulum torque

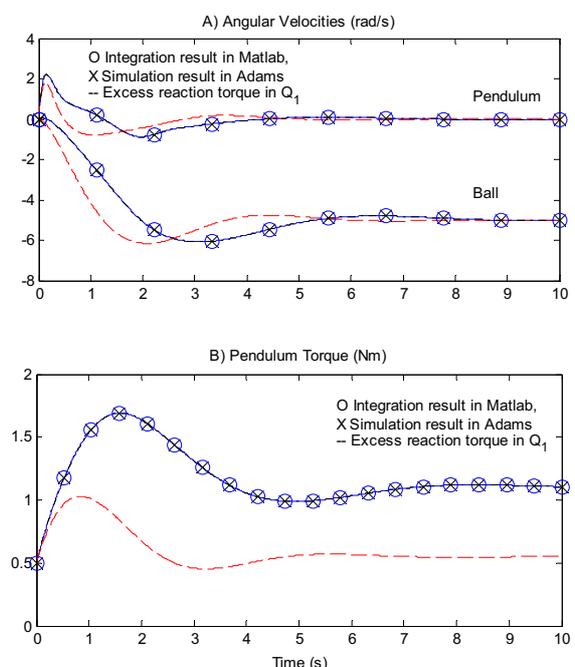


Figure 3: A) In a closed-loop, a PI-controller stabilizes the ball velocity. Excess reaction torque in Q_1 causes a different system response. B) Excess reaction torque causes error in the simulated pendulum driving torque

APPENDIX 1

THE COMPLETE DYNAMIC MODELS IN UNIFIED REPRESENTATION

Unless otherwise stated, the ball kinetic energy is: $T_1 = \frac{1}{2}M_1R^2\dot{\theta}_1^2 + \frac{1}{2}J_1\dot{\theta}_1^2$.

' τ ' presents the pendulum motor torque.

The given elements for A , B , C , D , G , and Q complete the configuration space presentation:

$$\begin{bmatrix} A11 & A12 \\ A21 & A22 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} B11 & B12 \\ B21 & B22 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1^2 \\ \dot{\theta}_2^2 \end{bmatrix} + [C11](\dot{\theta}_1\dot{\theta}_2) + \begin{bmatrix} D11 & D12 \\ D21 & D22 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} G11 \\ G21 \end{bmatrix} = \begin{bmatrix} Q1 \\ Q2 \end{bmatrix}$$

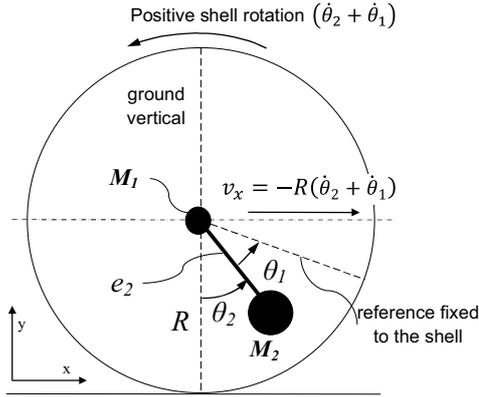


Figure A1: Absolute Pendulum Angle, Relative Ball Angle in Same Direction

A) Absolute Pendulum Angle, Relative Ball Angle in Same Direction

$$\begin{aligned} A11 &= J_1 + (M_1 + M_2)R^2 \\ A12, A21 &= J_1 + (M_1 + M_2)R^2 - M_2Re_2\cos(\theta_2) \\ A22 &= J_1 + J_2 + (M_1 + M_2)R^2 + M_2e_2^2 - 2M_2Re_2\cos(\theta_2) \\ B12, B22 &= M_2Re_2\sin(\theta_2) \\ D11 &= (c_1 + c_2) \\ D12, D21, D22 &= c_1 \\ G2 &= M_2ge_2\sin(\theta_2) \\ Q1 &= -\tau_1 \\ Q2 &= 0 \end{aligned}$$

Applicable References: (Koshiyama and Yamafuji 1992, 1993)

$$V_2 = -M_2ge_2\cos(\theta_2)$$

$$T_1 = \frac{1}{2}M_1R^2(\dot{\theta}_2 + \dot{\theta}_1)^2 + \frac{1}{2}J_1(\dot{\theta}_2 + \dot{\theta}_1)^2$$

$$T_2 = \frac{1}{2}M_2(R(\dot{\theta}_2 + \dot{\theta}_1) - e_2\dot{\theta}_2\cos(\theta_2))^2 + \frac{1}{2}M_2(e_2\dot{\theta}_2\sin(\theta_2))^2 + \frac{1}{2}J_2\dot{\theta}_2^2$$

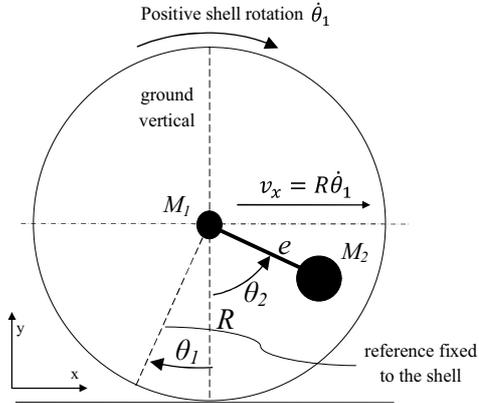


Figure A2: Absolute Ball Angle, Absolute Pendulum Angle in Opposite Direction

B) Absolute Ball Angle, Absolute Pendulum Angle in Opposite Direction

$$\begin{aligned} A11 &= J_1 + (M_1 + M_2)R^2 \\ A12, A21 &= M_2Re\cos(\theta_2) \\ A22 &= J_2 + M_2e^2 \\ B12 &= -M_2Resin(\theta_2) \\ D11 &= (c_1 + c_2) \\ D12, D21, D22 &= c_2 \\ G2 &= M_2e\sin(\theta_2)g \\ Q1 &= \tau \\ Q2 &= \tau \end{aligned}$$

Applicable References: (Yue et al. 2006; Liu et al. 2009; Zhang et al. 2009; Ghanbari et al. 2010; Liu et al. 2012; Yu et al. 2012a; Yu et al. 2012b; Yue and Liu 2012a; Yue and Liu 2012b; Mahboubi et al. 2013; Yue and Liu 2013;)

$$V_2 = -M_2ge\cos(\theta_2)$$

$$T_2 = \frac{1}{2}M_2((R\dot{\theta}_1 + e\dot{\theta}_2\cos(\theta_2))^2 + (e\dot{\theta}_2\sin(\theta_2))^2) + \frac{1}{2}J_2\dot{\theta}_2^2$$

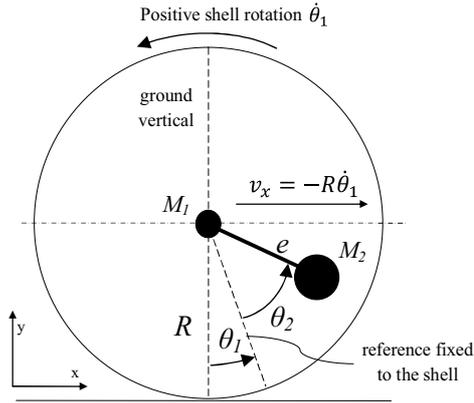


Figure A3: Absolute Ball Angle, Relative Pendulum Angle in Same Direction

$$V_2 = -M_2 g e \cos(\theta_1 + \theta_2)$$

$$T_2 = \frac{1}{2} M_2 (-R\dot{\theta}_1 + e(\dot{\theta}_1 + \dot{\theta}_2) \cos(\theta_1 + \theta_2))^2 + \frac{1}{2} M_2 (e(\dot{\theta}_1 + \dot{\theta}_2) \sin(\theta_1 + \theta_2))^2 + \frac{1}{2} J_2 (\dot{\theta}_1 + \dot{\theta}_2)^2$$

C) Absolute Ball Angle, Relative Pendulum Angle in Same Direction

$$A11 = J_1 + J_2 + M_2 e^2 + (M_1 + M_2) R^2 - 2M_2 R e \cos(\theta_1 + \theta_2)$$

$$A12, A21 = J_2 + M_2 e^2 - M_2 R e \cos(\theta_1 + \theta_2)$$

$$A22 = J_2 + M_2 e^2$$

$$B11, B12 = M_2 R e \sin(\theta_1 + \theta_2)$$

$$C11 = 2M_2 R e \sin(\theta_1 + \theta_2)$$

$$D11 = c_1$$

$$D22 = c_2$$

$$G1 = M_2 e \sin(\theta_1 + \theta_2) g$$

$$G2 = M_2 e \sin(\theta_1 + \theta_2) g$$

$$Q1 = 0$$

$$Q2 = \tau$$

Applicable References: (Nagai 2008, Kim et al. 2009)

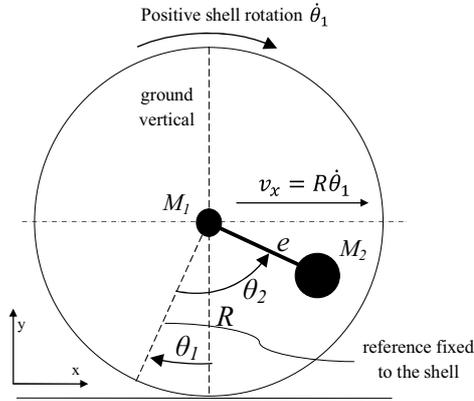


Figure A4: Absolute Ball Angle, Relative Pendulum Angle in Opposite Direction

$$V_2 = -M_2 g e \cos(\theta_2 - \theta_1)$$

$$T_2 = \frac{1}{2} M_2 ((R\dot{\theta}_1 + e(\dot{\theta}_2 - \dot{\theta}_1) \cos(\theta_2 - \theta_1))^2 + \frac{1}{2} M_2 (e(\dot{\theta}_2 - \dot{\theta}_1) \sin(\theta_2 - \theta_1))^2 + \frac{1}{2} J_2 (\dot{\theta}_2 - \dot{\theta}_1)^2$$

D) Absolute Ball Angle, Relative Pendulum Angle in Opposite Direction

$$A11 = J_1 + J_2 + M_2 e^2 + (M_1 + M_2) R^2 - 2M_2 R e \cos(\theta_1 - \theta_2)$$

$$A12, A21 = -J_2 - M_2 e^2 + M_2 R e \cos(\theta_1 - \theta_2)$$

$$A22 = J_2 + M_2 e^2$$

$$B11, B12 = M_2 R e \sin(\theta_1 - \theta_2)$$

$$C11 = -2M_2 R e \sin(\theta_1 - \theta_2)$$

$$D11 = c_1$$

$$D22 = c_2$$

$$G1 = M_2 e \sin(\theta_1 - \theta_2) g$$

$$G2 = -M_2 e \sin(\theta_1 - \theta_2) g$$

$$Q1 = 0$$

$$Q2 = \tau$$

Applicable References: (Kayacan et al. 2012a, 2012b, 2012c)

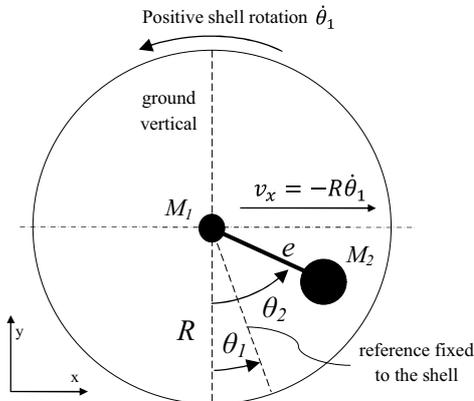


Figure A5: Absolute Ball Angle, Absolute Pendulum Angle in Same Direction

$$V_2 = -M_2 g e \cos(\theta_2)$$

$$T_2 = \frac{1}{2} M_2 ((-R\dot{\theta}_1 + e\dot{\theta}_2 \cos(\theta_2))^2 + (e\dot{\theta}_2 \sin(\theta_2))^2) + \frac{1}{2} J_2 \dot{\theta}_2^2$$

E) Absolute Ball Angle, Absolute Pendulum Angle in Same Direction

$$A11 = J_1 + (M_1 + M_2) R^2$$

$$A12, A21 = -M_2 R e \cos(\theta_2)$$

$$A22 = J_2 + M_2 e^2$$

$$B12 = M_2 R e \sin(\theta_2)$$

$$D11 = (c_1 + c_2)$$

$$D12, D21 = -c_2$$

$$D22 = c_2$$

$$G2 = M_2 e \sin(\theta_2) g$$

$$Q1 = -\tau$$

$$Q2 = \tau$$

Applicable Reference: (Cai et al. 2011)

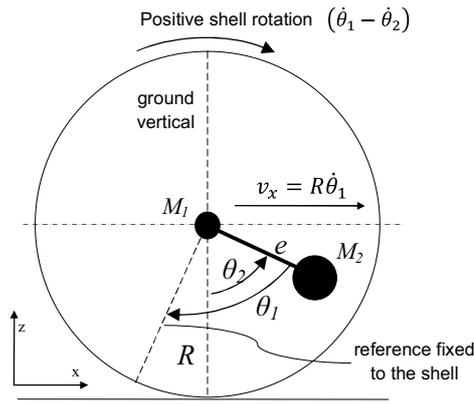


Figure A6: Absolute Pendulum Angle, Relative Ball Angle in Opposite Direction

F) Absolute Pendulum Angle, Relative Ball Angle in Opposite Direction

$$\begin{aligned}
 A11 &= J_1 + (M_1 + M_2)R^2 \\
 A12, A21 &= -J_1 - (M_1 + M_2)R^2 + M_2 R e \cos(\theta_2) \\
 A22 &= J_1 + J_2 + M_2 e^2 + (M_1 + M_2)R^2 - 2M_2 R e \cos(\theta_2) \\
 B12 &= -M_2 R e \sin(\theta_2) \\
 B22 &= M_2 R e \sin(\theta_2) \\
 D11 &= (c_1 + c_2) \\
 D12, D21 &= -c_1 \\
 D22 &= c_1 \\
 G2 &= M_2 e \sin(\theta_2) g \\
 Q1 &= \tau \\
 Q2 &= 0
 \end{aligned}$$

Applicable Reference: (Kamaldar et al. 2011)

$$V_2 = -M_2 g e \cos(\theta_2)$$

$$T_1 = \frac{1}{2} M_1 (R(\dot{\theta}_1 - \dot{\theta}_2))^2 + \frac{1}{2} J_1 (\dot{\theta}_1 - \dot{\theta}_2)^2$$

$$T_2 = \frac{1}{2} M_2 ((R(\dot{\theta}_1 - \dot{\theta}_2) + e\dot{\theta}_2 \cos(\theta_2))^2) + \frac{M_2}{2} (e\dot{\theta}_2 \sin(\theta_2))^2 + \frac{J_2}{2} (\dot{\theta}_2)^2$$

REFERENCES

- Balandin, D.V.; Komarov, M.A. and Osipov, G.V. 2013. "A motion control for a spherical robot with pendulum drive." *Journal of Computer and Systems Sciences International* 52, No.4, 650-663.
- Bicchi, A.; Balluchi, A.; Prattichizzo, D. and Gorelli, A. 1997. "Introducing the "SPHERICLE": an experimental testbed for research and teaching in nonholonomy." In *Proceedings of IEEE International Conference on Robotics and Automation* (Albuquerque, NM, April 1997). 2620-2625.
- Bicchi, A.; Prattichizzo, D. and Sastry, S.S. 1995. "Planning motions of rolling surfaces." In *Proceedings of IEEE Conference on Decision and Control* (New Orleans, Dec. 1995). 2812-2817.
- Bruhn, F.C.; Pauly, K. and Kaznov, V. 2005. "Extremely low mass spherical rovers for extreme environments and planetary exploration enabled with MEMS." *European Space Agency, (Special Publication) ESA SP 603*, 347-354.
- Cai, Y.; Zhan, Q. and Xi, X. 2011. "Neural Network Control for the Linear Motion of a Spherical Mobile Robot." *International Journal Of Advanced Robotic Systems* 8, No.4, 79-87.
- Camicia, C.; Conticelli, F. and Bicchi, A. 2000. "Nonholonomic kinematics and dynamics of the Sphericle." In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (Takamatsu, Japan, Nov. 2000). 805-810.
- Das, T. and Mukherjee, R. 2004. "Exponential stabilization of the rolling sphere." *Automatica*, 2004, Vol.40(11), pp.1877-1889 40, No.11, 1877-1889.
- Das, T. and Mukherjee, R. 2006. "Reconfiguration of a rolling sphere: A problem in evolute-involute geometry." *ASME Journal of Applied Mechanics* 73, No.4, 590-597.
- Gajbhiye, S. and Banavar, R.N. 2012. "The euler-poincaré equations for a spherical robot actuated by a pendulum." In *Proceedings of 4th IFAC Workshop on Lagrangian and Hamiltonian Methods for Non Linear Control* (Bertinoro, Aug. 2012). 42-47.
- Ghanbari, A.; Mahboubi, S. and Fakhrabadi, M.M.S. 2010. "Design, dynamic modeling and simulation of a spherical mobile robot with a novel motion mechanism." In *Proceedings of IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications* (Qingdao, Shan Dong, July 2010). 434-439.
- Goldstein, H.; Poole, C. and Safko, J. 2002. *Classical mechanics*. 3rd ed. Addison-Wesley, San Francisco (CA).
- Halme, A.; Schonberg, T. and Wang, Y. 1996. "Motion control of a spherical mobile robot." In *Proceedings of IEEE International Workshop on Advanced Motion Control* (Mie, 18-21 Mar.). 259-264.
- Jia, Q.; Sun, H. and Liu, D. 2008. "Analysis of Actuation for a Spherical Robot." In *Proceedings of IEEE Conference on Robotics, Automation and Mechatronics* (Chengdu, China, 21-24 Sept.). 266-271.
- Jia, Q.; Zheng, Y.; Sun, H.; Cao, H. and Li, H. 2009. "Motion control of a novel spherical robot equipped with a flywheel." In *Proceedings of IEEE International Conference on Information and Automation* (Zhuhai, Macau, June 2009). 893-898.
- Joshi, V.A. and Banavar, R.N. 2009. "Motion analysis of a spherical mobile robot." *Robotica* 27, No.3, 343-353.
- Joshi, V.A.; Banavar, R.N. and Hippalgaonkar, R. 2007. "Design, modeling and controllability of a spherical mobile robot." In *Proceedings of National Conference on Mechanisms and Machines* (Bangalore, India, Dec. 2007). 135-140.
- Joshi, V.A.; Banavar, R.N. and Hippalgaonkar, R. 2010. "Design and analysis of a spherical mobile robot." *Mechanism and Machine Theory* 45, No.2, 130-136.
- Jurdjevic, V. 1993. "The geometry of the plate-ball problem." *Archive for Rational Mechanics and Analysis* 124, No.4, 305-328.
- Kamaldar, M.; Mahjoob, M.J.; Haeri Yazdi, M.; Vahid-Alizadeh, H. and Ahmadizadeh, S. 2011. "A control

- synthesis for reducing lateral oscillations of a spherical robot." In *Proceedings of IEEE International Conference on Mechatronics* (Istanbul, Apr. 2011). 546-551.
- Karimpour, H.; Keshmiri, M. and Mahzoon, M. 2012. "Stabilization of an autonomous rolling sphere navigating in a labyrinth arena: A geometric mechanics perspective." *Systems & Control Letters* 61, No.4, 495-505.
- Kayacan, E.; Bayraktaroglu, Z.Y. and Saeys, W. 2012a. "Modeling and control of a spherical rolling robot: a decoupled dynamics approach." *Robotica* 30, No.4, 671-680.
- Kayacan, E.; Kayacan, E.; Ramon, H. and Saeys, W. 2012b. "Velocity Control of a Spherical Rolling Robot Using a Grey-PID Type Fuzzy Controller With an Adaptive Step Size." In *Proceedings of 10th International IFAC Symposium on Robot Control* (Dubrovnik, Croatia, September 5 - 7). 863-868.
- Kayacan, E.; Kayacan, E.; Ramon, H. and Saeys, W. 2012c. "Adaptive Neuro-Fuzzy Control of a Spherical Rolling Robot Using Sliding-Mode-Control-Theory-Based Online Learning Algorithm." *Transactions on Systems Man and Cybernetics Part C* 43, No.1, 170-179.
- Kaznov, V. and Seeman, M. 2010. "Outdoor navigation with a spherical amphibious robot." In *Proceedings of Intelligent Robots and Systems* (Taipei, Oct. 2010). 5113-5118.
- Kim, J.; Kwon, H. and Lee, J. 2009. "A rolling robot: Design and implementation." In *Proceedings of 7th Asian Control Conference* (Hong Kong, Aug. 2009). 1474-1479.
- Koshiyama, A. and Yamafuji, K. 1992. "Development and motion control of the all direction steering-type mobile robot (1st report, a concept of spherical shaped robot, roll and running control)." *Transactions of the Japan Society of Mechanical Engineers, Series C* (In Japanese) 58, No.548, 1128-1136.
- Koshiyama, A. and Yamafuji, K. 1993. "Design and control of an all-direction steering type mobile robot." *International Journal Of Robotics Research* 12, No.5, 411-419.
- Li, Z. and Canny, J. 1990. "Motion of two rigid bodies with rolling constraint." *IEEE Transactions on Robotics and Automation* 6, No.1, 62-72.
- Liu, B.; Yue, M. and Liu, R. 2012. "Motion control of an underactuated spherical robot: A hierarchical sliding-mode approach with disturbance estimation." In *Proceedings of IEEE International Conference on Mechatronics and Automation* (Chengdu, Aug. 2012). 1804-1809.
- Liu, D. and Sun, H. 2010. "Nonlinear sliding-mode control for motion of a spherical robot." In *Proceedings of 29th Chinese Control Conference* (Beijing, 29-31 July). 3244-3249.
- Liu, D.; Sun, H. and Jia, Q. 2008. "Stabilization and path following of a spherical robot." In *Proceedings of IEEE International Conference on Robotics, Automation and Mechatronics* (Chengdu, Sept. 2008). 676-682.
- Liu, D.; Sun, H. and Jia, Q. 2009. "A Family of Spherical Mobile Robot: Driving Ahead Motion Control by Feedback Linearization." In *Proceedings of 2nd International Symposium on Systems and Control in Aerospace and Astronautics* (Shenzhen, Dec. 2008). 1-6.
- Liu, D.; Sun, H.; Jia, Q. and Wang, L. 2008. "Motion control of a spherical mobile robot by feedback linearization." In *Proceedings of the World Congress on Intelligent Control and Automation* (Chongqing, June 2008). 965-970.
- Mahboubi, S.; Seyyed Fakhraabadi, M.M. and Ghanbari, A. 2013. "Design and implementation of a novel spherical mobile robot." *Journal of Intelligent and Robotic Systems: Theory and Applications* 71, No.1, 43-64.
- Michaud, F. and Caron, S. 2002. "Roball, the rolling robot." *Autonomous Robots* 12, No.2, 211-222.
- Morinaga, A.; Svinin, M. and Yamamoto, M. 2012. "On the motion planning problem for a spherical rolling robot driven by two rotors." In *Proceedings of IEEE/SICE International Symposium on System Integration* (Fukuoka, Dec. 2012). IEEE, 704-709.
- Mukherjee, R.; Minor, M.A. and Pukrushpan, J.T. 1999. "Simple motion planning strategies for spherobot: a spherical mobile robot." In *Proceedings of the 38th IEEE Conference on Decision and Control* (AZ, 1999). 2132-2137.
- Mukherjee, R.; Minor, M.A. and Pukrushpan, J.T. 2002. "Motion planning for a spherical mobile robot: Revisiting the classical ball-plate problem." *ASME Journal of Dynamic Systems, Measurement, and Control* 124, No.4, 502-511.
- Nagai, M. 2008. *Control System of a Ball-shaped Robot*. Master's Thesis, Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in technology, Department of Automation and Systems Technology, Helsinki University of Technology.
- Sang, S.J.; Shen, D.; Zhao, J.C.; Hu, J.Y. and An, Q. 2011. "Analysis and Simulation of a Spherical Robot." *Advanced Materials Research* 171 - 172, 748-751.
- Svinin, M. and Hosoe, S. 2008. "Motion Planning Algorithms for a Rolling Sphere With Limited Contact Area." *IEEE Transactions on Robotics* 24, No.3, 612-625.
- Svinin, M.; Morinaga, A. and Yamamoto, M. 2012a. "An analysis of the motion planning problem for a spherical rolling robot driven by internal rotors." In *Proceedings of IEEE International Conference on Intelligent Robots and Systems* (Vilamoura, Oct. 2012). IEEE, 414-419.
- Svinin, M.; Morinaga, A. and Yamamoto, M. 2012b. "On the dynamic model and motion planning for a class of spherical rolling robots." In *Proceedings of IEEE International Conference on Robotics and Automation* (Saint Paul, May 2012). 3226-3231.
- Symon, K.R. 1960. *Mechanics*. 2nd ed. Addison-Wesley, Reading, MA.
- Ylikorpi, T. and Suomela, J. 2007. "Ball-shaped Robots." In *Climbing & Walking Robots, Towards New Applications*, Z. Houxiang Zhang (Ed.). InTech, Vienna, 235-256.
- Yoon, J.C.; Ahn, S.S. and Lee, Y.J. 2011. "Spherical robot with new type of two-pendulum driving mechanism." In *Proceedings of INES 2011 - 15th International Conference on Intelligent Engineering Systems* (Poprad, Slovakia, June 23-25). IEEE, 275-279.
- Yu, T.; Sun, H. and Zhang, Y. 2011. "Dynamic analysis of a spherical mobile robot in rough terrains." In *Proceedings of SPIE 8044, Sensors and Systems for Space Applications IV, 80440V* (Orlando, Apr. 2011).
- Yu, T.; Sun, H.; Jia, Q.; Zhang, Y. and Zhao, W. 2012a. "Sliding mode control of pendulum-driven spherical robots." *Advanced Materials Research* 591-593, 1519-1522.
- Yu, T.; Sun, H.; Zhang, Y. and Zhao, W. 2012b. "Control and stabilization of a pendulum-driven spherical mobile robot on an inclined plane." In *Proceedings of 11th International Symposium on Artificial Intelligence, Robotics and Automation in Space* (Turin, Italy, 4 - 6 Sept.).
- Yue, M.; Deng, Z.; Yu, X. and Yu, W. 2006. "Introducing HIT Spherical Robot: Dynamic Modeling and Analysis Based on Decoupled Subsystem." In *Proceedings of IEEE International Conference on Robotics and Biomimetics* (Kunming, China, December 17-20). 181-186.

- Yue, M. and Liu, B. 2012a. "Disturbance adaptive control for an underactuated spherical robot based on hierarchical sliding-mode technology." In *Proceedings of Chinese Control Conference* (Hefei, July 2012). 4787-4791.
- Yue, M. and Liu, B. 2012b. "Design of adaptive sliding mode control for spherical robot based on MR fluid actuator." *Journal of Vibroengineering* 14, No.1, 196-204.
- Yue, M. and Liu, B. 2013. "Adaptive control of an underactuated spherical robot with a dynamic stable equilibrium point using hierarchical sliding mode approach." *International Journal of Adaptive Control and Signal Processing*.
- Zhan, Q.; Cai, Y. and Yan, C. 2011. "Design, analysis and experiments of an omni-directional spherical robot." In *Proceedings of IEEE International Conference on Robotics and Automation* (Shanghai, May 2011). IEEE, 4921-4926.
- Zhan, Q.; Zhou, T.; Chen, M. and Cai, S. 2006. "Dynamic Trajectory Planning of a Spherical Mobile Robot." In *Proceedings of IEEE Conference on Robotics, Automation and Mechatronics* (Bangkok, June 2006). 1-6.
- Zhang, Q.; Jia, Q.; Sun, H. and Gong, Z. 2009. "Application of a Genetic Algorithm-Based PI Controller in a Spherical Robot." In *Proceedings of IEEE International Conference on Control and Automation* (Christchurch, Dec. 2009). 180-184.
- Zheng, M.; Zhan, Q.; Liu, J. and Cai, Y. 2011. "Control of a Spherical Robot: Path Following Based on Nonholonomic Kinematics and Dynamics." *Chinese Journal of Aeronautics* 24, No.3, 337-345.
- Zheng, Y.L. 2011. "Dynamic Analysis and Control Experiment of a Novel Mobile Robot." *Applied Mechanics and Materials* 58 - 60, 1577-1582.
- Zhuang, W.; Liu, X.; Fang, C. and Sun, H. 2008. "Dynamic modeling of a spherical robot with arms by using Kane's method." In *Proceedings of 4th International Conference on Natural Computation* (Jinan, Oct. 2008). 373-377.

AUTHOR BIOGRAPHIES

TOMI YLIKORPI went to Helsinki University of Technology and obtained his Master's degree in 1994 in Mechanical Engineering, Majoring in Mechatronics, having Minor in Space Technology. After graduation, he worked for 7 years in development of space instruments in Finland and in Italy. After returning to Finland, he obtained his Licentiate's degree in Automation Technology in 2008 and continues working at the Aalto University with space-related projects and designing and modelling of mechanical systems. His e-mail address is: TOMI.YLIKORPI@AALTO.FI and his web-page with further information can be found at [HTTP://AUTSYS.AALTO.FI/EN/TOMIYLIKORPI](http://AUTSYS.AALTO.FI/EN/TOMIYLIKORPI)

PEKKA FORSMAN received his PhD in Automation technology from Helsinki University of Technology in 2001. He is currently university lecturer at Aalto University. His research interests include field and service robotics, human-robot interfaces as well as localization and navigation methods. His e-mail address is: PEKKA.FORSMAN@AALTO.FI

AARNE HALME started his career in 1966 at Helsinki University of Technology. Since then he has acted as an Associate Professor, and as a Professor and Head of the Control and Systems Engineering Laboratory in Tampere University and Oulu University. From 1985 until lately he has been the Professor and Head of the Automation Technology Laboratory at Helsinki University of Technology. He is one of the very first pioneers in development, analysis and applications of ball-shaped rolling robots. His e-mail address is: AARNE.HALME@AALTO.FI and web-page is available at [HTTP://AUTOMATION.TKK.FI/FILES/AHALME/](http://AUTOMATION.TKK.FI/FILES/AHALME/)

JARI SAARINEN received his M.Sc. degree in 2002 and his PhD in Automation technology in 2009 from Helsinki University of Technology. Since graduation he has acted as a senior researcher in Center of Excellence in Generic Intelligent Machines (GIM) at Aalto University, and as researcher in Centre for Applied Autonomous Sensor Systems at Örebro University. His research interests include long-term autonomy, 3D perception, localization and mapping. He is currently chief executive officer in GIM Ltd., pursuing for real world robotic applications. His e-mail address is: JARI.P.SAARINEN@GMAIL.COM

INTEGRATING SIMULATION WITH ROBOTIC LEARNING FROM DEMONSTRATION

Anat Hershkovitz Cohen and Sigal Berman
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
POB 653, Beer-Sheva, Israel
E-mail: anhe@post.bgu.ac.il

KEYWORDS

Dynamic Motion Primitives, Reinforcement Learning, Simulation, Robotics.

ABSTRACT

Robots that co-habitat an environment with humans, e.g., in a domestic or an agricultural environment, must be capable of learning task related information from people who are not skilled in robotics. Learning from demonstration (LfD) offers a natural way for such communication. Learning motion primitives based on the demonstrated trajectories facilitate robustness to dynamic changes in the environment and task. Yet since the robot and human operator typically differ, a phase of autonomous learning is needed for optimizing the robotic motion. Autonomous learning using the physical hardware is costly and time consuming. Thus finding ways to minimize this learning time is of importance. In the current paper we investigate the contribution of integrating an intermediate stage of learning using simulation, after LfD and before learning using robotic hardware. We use dynamic motion primitives for motion planning, and optimize their learned parameters using the PI^2 algorithm which is based on reinforcement learning. We implemented the method for reach-to-grasp motion for harvesting an artificial apple. Our results show learning using simulation drastically improves the robotic paths and that for reach-to-grasp motion such a stage may eliminate the need for learning using physical hardware. Future research will test the method for motion that requires interaction with the environment.

INTRODUCTION

Robots and humans are starting to share common workspaces, where the robot operators are not necessarily skilled in robotics. Such shared workplaces are appearing in industry, agriculture, the medical establishment, and domestic environments. Methods for natural interaction between the human and the robot are a crucial component in such scenarios. It is additionally advantageous for such robotic co-workers to have motion characteristics that bear similarities to human

motion as this can aid robot acceptance, cooperation, and safety.

In the agricultural environment, robot penetration is expected to increase considerably in the current decade (Forge and Blackman 2011). Selective harvesting of high-value crops, e.g., apples and peppers, is an appealing task for automation using robotic systems (Foglia and Reina 2006). Selective harvesting is seasonal, tedious, only ripe fruit should be picked, and fruit position and orientation on the branch are random. Selective harvesting requires dexterous manipulation capabilities so plant, fruit, and future fruit growth are not damaged (Allota et al. 1990).

In Learning from Demonstration (LfD), also termed programming by demonstration (PbD) or imitation learning, the robot learns from demonstrations of a teacher (Mataric 2007; Koenig et al. 2010). For teaching the robot what paths to follow and how, LfD can replace traditional robot programming methods which are tedious and require specialized knowledge (Argall et al. 2009). LfD can ease the deployment of robots for selective harvesting by alleviating the effort required from the farmer while enhancing the similarity of the robotic and human motion.

Motion primitives are simple motion elements that can be concatenated serially or in parallel. Many studies show that voluntary motion in both vertebrates and invertebrates is composed of such elements (Flash and Hochner 2005). Demonstrated trajectories can be used for learning motion primitives rather than just the direct path representation, to facilitate learning generalization and thus robustness to dynamic changes in the environment and task during execution. Dynamic Movement Primitives (DMPs) encode motion control parameters using nonlinear differential equations, where equation parameters can be learned from demonstration (Ijspeert et al. 2002; Tamosiunaite et al. 2011; Kulvicius et al. 2012). The control equations of a DMP are:

$$\frac{1}{\tau} \dot{v} = \alpha_v (\beta_v (g - p) - v) + f(s) \quad (1)$$

$$\frac{1}{\tau} \dot{p} = v \quad (2)$$

where p is the position and v the velocity of the system and g is the desired goal of the trajectory. The parameters α_v and β_v are constants and τ is a time scaling constant. Equation (1), termed the controller equation, contains the nonlinear component f which is defined based on the demonstrated movement as a weighted sum of Gaussian functions:

$$f(s) = \frac{\sum_i \psi_i(s) w_i}{\sum_i \psi_i(s)} s \quad (3)$$

$$\frac{1}{\tau} \dot{s} = -\alpha s \quad (4)$$

Where ψ_i are the Gaussian functions and w_i are their weights, f depends explicitly on the phase variable, s , and α is a predefined constant.

The phase variable eliminates the time dependency and facilitates changing the movement duration using the scaling parameter, τ , without changing the trajectory. Equation (4) is termed the ‘canonical system’. As s approaches one at the beginning of the trajectory and zero at the end of the movement causing f to vanish from the controller equation (equation 1) which then linearly converges to the goal, g (Nakanishi et al. 2004; Ijspeert et al. 2003; Pastor et al. 2009). DMPs facilitate learning dynamic trajectory characteristics while allowing adaptation to change during run-time (Hoffman et al. 2009). These characteristics are both crucial for robots in dynamic environments, such as the agricultural environment.

When there is a large difference between the demonstrating agent, e.g., the human, and the learning agent, e.g., the robot, several researchers have suggested a phase of autonomous training following the LfD phase. The purpose of the autonomous learning phase is to adapt the learned capabilities to the actual characteristics of the performing agent. Several Reinforcement learning (RL) methods have been suggested for the autonomous learning phase (Kormushev et al. 2010). RL is a commonly used method for learning where task dynamics are difficult to model (Sutton and Barto, 1998). Yet RL does not scale well to high dimensional search spaces in which it suffers from convergence problems. Using the parameters learned in the initial LfD phase as the starting point for RL-based autonomous learning can expedite the convergence, and help avoid local optima.

Policy Improvement with Path Integrals (PI²) is a variant of RL that has been found to outperform other RL-based learning algorithms for DMP parameter optimization (Kappen 2007; Broek et al. 2008). PI² is an RL method based on stochastic optimal control where cost is minimized rather than reward maximized. PI² can scale to high dimensional problems rendering it applicable for physical robotic systems with many degrees of freedom (DOF) (Buchli et al. 2011;

Theodorou et al. 2010). The learning procedure of PI² is organized in epochs, i.e., several trials (roll-out) are run in which exploration noise is added to the weights of the Gaussian functions of the DMP. The cost of each roll-out is evaluated according to the cost value function. The weights are then updated based on the cost of all the roll-outs in the epoch. An additional advantage of PI² is that it does not require parameter tuning apart from the exploration noise (Tamosiunaite et al. 2011). Recent papers present practical implementations of the PI² algorithm including a robot dog that jumps across a gap (Pastor et al. 2013; Theodorou et al. 2010), a robotic arm which pours liquid (Tamosiunaite et al. 2011), a robotic arm that opens a door using its handle and a robotic arm equipped with a three-fingered hand and a force-torque sensor at the wrist, that picks a pen from table (Kalakrishnan et al. 2011).

The cost of learning using a physical system is generally high, thus it is important to minimize the time required for such a learning stage. In the current study we explore the integration of an intermediate stage of learning using a simulation of the physical system and its contribution to improvement in performance. The study aims to answer whether integration of an autonomous learning stage using a simulation model of the system can reduce the need for learning using the physical system. The rest of this paper is organized as follows: The method section presents the conducted experiments along with a description of the implementation of the different learning phases. It is followed by a Results section which presents the results of the experiment along with a discussion of their implications. The results section is followed by a Conclusions and Future Research section.

METHOD

Environment and task

A reach-to-grasp task was chosen for examination of the contribution of the different learning phases. Apples are harvested by a wrist motion that applies shear force against their peduncle. It is important that the stem remains connected to the apple after harvesting otherwise the apple cannot be marketed as fresh produce. Reaching the apple in a pose (position and orientation), that will enable the motion required for detaching it, is crucial for successful harvesting.

The task was conducted by both the human demonstrators and the robotic manipulator in the telerobotics laboratory, at the Ben-Gurion University of the Negev (Figure 1). An artificial apple tree was located in the center of the laboratory. Apples are connected to the artificial tree such that pulling them from the branch detaches them from it.

Learning

The learning process was divided into three sequential phases (Figure 2). First, a set of movements for each

task was recorded from human demonstrators (demonstration phase). A DMP was created for each axis based on the demonstrated movements using LfD (LfD phase) and then optimized first in simulation and then using physical hardware in the autonomous training phase.

Demonstration phase

Three demonstrators participated in the demonstration phase. Each demonstrator stood with both arms alongside the body in front of the tree, at an arms-length from it. A six DOF magnetic motion sensor (FASTRAK™, Pulhamus) was attached to the demonstrator's wrist. Each demonstrator executed two harvesting movements, where the harvesting hand pose was once to the side and once in front of the apple. Each harvesting movement was divided into two sub-movements: a reach-to-grasp movement that ended when the demonstrator grasped the apple and a detachment movement. The demonstrators were requested to pause shortly between the two sub-movements. After each harvesting movement the apple was re-inserted into its place on the tree.



Figure 1: Reach-to-grasp of an artificial apple in the telerobotics laboratory. Top: The human demonstrator, Bottom: The Motoman UP6 robot.

LfD phase

The demonstrated trajectories were transformed to robot-base coordinates. The recorded movement was very noisy thus it was smoothed with a 7th degree polynomial and resampled at constant intervals such that each trajectory was represented by N=24 points. Three DMPs were created, one for each main axis in robot-base coordinates. The nonlinear movement component was extracted from the demonstrated movement and approximated using non-linear regression by a weighted sum of 10 Gaussian functions. The start and goal points of each axis were defined based on robot position and a trajectory was created for each axis using the computed DMP.

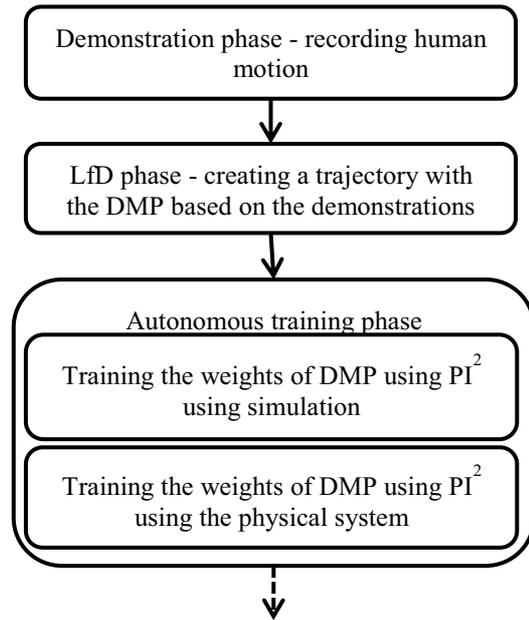


Figure 2: The learning process

Autonomous training using simulation

The PI² algorithm was implemented using MatLab (MathWorks, USA). The algorithm was run on an Intel® Core™ i7-2600 3.40 Ghz processor, with 4 GB RAM, and windows 7 Enterprise, 32 bit operation system.

Two cost functions were defined for creating a smooth robotic motion: minimum squared acceleration (MSA) and minimum acceleration change (MCA),

$$MSA_j = \frac{\sum_{i=1}^N a_j^2(i)}{N * a_{\max}^2} + \frac{(p_j(N) - g_j)^2}{\Delta p g_{\max}^2} + \frac{\sum_{i=1}^N (p_j(i) - d_j(i))^2}{N * \Delta p d_{\max}^2} \quad (5)$$

Where j indicates the DMP axis, For the trajectory created by the DMP, $a(i)$ is the acceleration and $p(i)$ is the position at sample i , $d(i)$ the position of the demonstrated movement. The values of a_{max} , Δpg_{max} and Δpd_{max} are the overall maxima of acceleration, difference between final position and goal over all the roll outs, and difference between position and demonstrated position respectively.

$$MCA_j = \frac{\sum_{i=2}^N \left(\frac{a_j(i) - a_j(i-1)}{T/(N-1)} \right)^2}{N * \Delta a_{max}^2} + \frac{(p_j(N) - g_j)^2}{\Delta pg_{max}^2} + \frac{\sum_{i=1}^N (p_j(i) - d_j(i))^2}{N * \Delta pd_{max}^2} \quad (6)$$

Where Δa_{max} is the overall maxima of the change in accelerations, and T is the movement duration.

Both cost functions include three parameters, two of which are similar: the distance to the goal and the similarity to the demonstrated movement. The distance to the goal is measured by the distance between the actual final configuration and the planned goal. The similarity to the demonstrated movement is measured by the sum of the squared errors between the performed trajectory and the demonstrated trajectory. MSA additionally includes the acceleration amplitude measured by the sum of squared accelerations and MCA includes the change in accelerations measured by the sum of squared changes in acceleration. The parameters were chosen such that the final trajectory will be similar to the demonstrated path, converge to the goal, and will be smooth. All the parameters are normalized and equally weighted. The robotic motion is simulated based on classical motion equations.

The initial inputs to the simulation-learning phase were the DMPs computed based on the demonstrations. Ten roll-outs were executed in each iteration of the PI² algorithm. For each roll-out a vector of random numbers was sampled from the standard normal distribution. The random values were multiplied by a predefined constant. The constant was set by trial and error to be 20. The rate of change when using the MSA cost function converged to zero after approximately 5000 iterations and thus the number of iterations was set to 5000 for all trials.

Autonomous training using the physical system

The robot control software was programed using Microsoft C# and MotoCom robotic communication library for C++ (Motoman, Japan). The communication between the robot controller and the computer was established through a serial RS232 link. The robot program was written using Infrom (Motoman, Japan) for an XRC Motoman controller.

The inputs to the training phase using the physical system were the DMPs after the simulation phase. Iterations of ten roll-outs were executed using the robot. The roll-outs were produced using MatLab. The noise added to the weights was computed as in the simulation runs. A trajectory based on each roll-out was sent to the robot for execution. The trajectory of the robot contained 24 via positions and took about 46 sec to complete. During the execution of the trajectory, the robot's current position was sampled at constant time intervals of 2 sec. After consecutively executing the ten roll-outs, these positions were used to compute the cost function and to accordingly update the DMP weights.

Analysis

For all recorded trajectories three DMPs were computed based on LfD and then optimized with 5000 iterations of the PI² algorithm using simulation. Average improvement in all the parameters of each cost function (MSA and MCA) was computed based on the values of the simulated trajectory.

For one of the trajectories the DMPs after the simulation-training were used as input for training using the physical system. Using the physical system the PI² algorithm was run for two iterations using the MSA cost function. For this trajectory, improvement was measured based on the actual trajectory performed by the robot. A trajectory was formed based on the three final DMPs computed after each phase (initial DMP computed based on LfD, training using simulation, training using the physical system) and sent to the robot for execution. The values of the cost function parameters were computed based on the trajectories executed by the robot.

RESULTS

Training using simulation

A typical learning curve is presented in Figure 3. This curve depicts the value of the cost function as a function of the executed iterations. From the graph it is apparent that the value of the cost function converges after 5000 iterations. The average run-time of 5000 iterations was 17 minutes.

The improvement in all parameters based on simulated trajectories after 5000 iterations of the PI² algorithm using simulation is presented in Table 1. The parameters should be minimized, thus, negative percentages present a decrease in the parameter (and thus improvement). Overall both cost functions (MSA and MCA) improved (reduced) by more than 50%. The final distance to the goal considerably decreased after the simulation-training phase using both MSA and MCA cost functions. The acceleration values (MSA) and the change in accelerations also considerably decreased. On the other-hand the distance to the demonstrated

trajectory increased as there is a tradeoff between the parameters.

Typical trajectories are depicted in Figure 4. The trajectories were produced for the human demonstration and based on the robot's movement after the different phases: the initial trajectory based on the DMPs after the LfD process and the trajectory produced by DMPs after 5000 iterations of the PI² algorithm using simulation. From the figure we can see that the trajectory produced by the initial DMPs is very similar to the demonstrated trajectory. The trajectory after the simulation has indeed changed yet, as required, it is still reminiscent of the original trajectory.

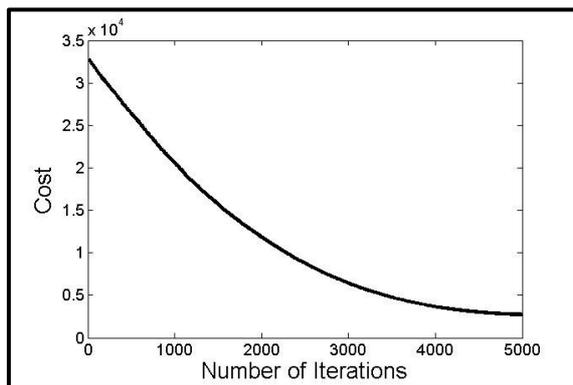


Figure 3: The cost as a function of the number of iterations using the MSA cost function

Table 1: Improvement based on simulated trajectories after training using simulation

	Distance to goal % (S.E.)	Distance to Demo % (S.E.)	Acceleration (change/value) % (S.E.)	Total % (S.E.)
MSA	-98 (0.05)	47.9 (0.6)	-21.9 (0.1)	-52.3 (0.2)
MCA	-85.5 (0.2)	166.6 (2.1)	-51.9 (0.3)	-53.1 (0.3)

S.E. – standard error

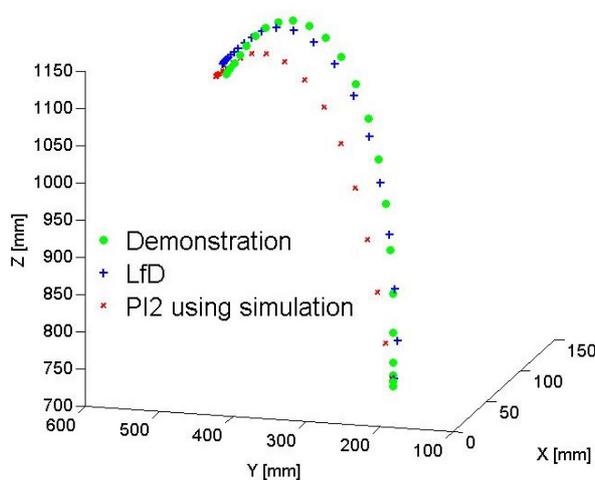


Figure 4: Reach-to-grasp trajectories: demonstrated, executed by the robot based on the initial DMPs learned from demonstration and after learning using simulation

Training using the physical system

For the selected trajectory and both cost functions the improvement based on actual robot trajectories (a single trajectory was tested for each cost function) in all parameters after 5000 iterations of the PI² algorithm in simulation is presented in Table 2.

Total cost is reduced using both cost functions. The final distance to the goal was reduced for both cost functions and the acceleration values (MSA) and the change in accelerations also considerably decreased. For MCA even the distance to the demonstrated trajectory decreased while for MSA this distance increased.

The run-time of two iterations of the PI² algorithm using the robot was 25 min. These iterations did not change the values of the cost function or its parameters.

Table 2: Improvement based on a single robot trajectory after training using simulation

	Distance to goal %	Distance to Demo %	Acceleration (change/value) %	Total %
MSA	-73.0	-14.2	-4	-69.5
MCA	-61.4	3.4	-20.6	-27.4

CONCLUSIONS AND FUTURE RESEARCH

For the reach-to-grasp motion in an apple harvesting task, the parameters of DMPs were learned based on human demonstration, and then adapted with the PI² algorithm using simulation and using hardware. The simulation-based training, which greatly improved the motion parameters, was considerably faster than hardware-based training. Training using hardware following the training using simulation did not additionally improve motion parameters.

Future work will examine integration of training using simulation in additional tasks and scenarios that require interaction with the environment, e.g., the apple detachment motion. We expect that training using hardware to be of importance in such tasks in which interaction dynamics are of importance.

ACKNOWLEDGMENT

The research is supported by the European Commission in the 7th Framework Programme (CROPS GA no 246252). The authors thank Noam Peles, Gil Baron, Nissim Abuhatzira and Josef Zahavi for their assistance in the hardware implementation.

REFERENCES

- Allota, B., Buttazzo, G., Dario, P., and Ouaqlia, F. 1990. "A Force/Torque Sensor-Based Technique for Robot Harvesting of Fruits and Vegetables". In *Intelligent Robots and Systems '90 'Towards a New Frontier of Applications'* (Ibaraki, July 3-6). IEEE, 231-235.
- Argall, B. D., Chernova, S., Veloso, M., and Browning, B. 2009. "A Survey of Robot Learning from Demonstration."

- Robotics and Autonomous Systems* 57, No. 5 (May), 469-483.
- Bekey, G.A. 2005. *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge, Mass. MIT Press.
- Broek, B. v. d., Wiegerinck, W., and Kappen, B. 2008. "Graphical Model Inference in Optimal Control of Stochastic Multi-Agent Systems." *Journal of Artificial Intelligence Research* 32, No.1 (Oct), 95-122.
- Buchli, J., Theodorou, E., Stulp, F., and Schaal, S. 2011. "Learning Variable Impedance Control." *The International Journal of Robotics Research* 30, No. 7 (June), 820-833.
- Flash, T., Hochner, B. 2005. "Motor Primitives in Vertebrates and Invertebrates," *Current Opinion in Neurobiology* 15, No. 6 (Dec), 660-666.
- Foglia, M.M. and Reina, G. 2006. "Agricultural Robot for Radicchio Harvesting." *Journal of Field Robotics* 23, No. 6 (July), 363-377.
- Forge, S. and Blackman, C. 2010. "A Helping Hand for Europe: The competitive outlook for EU robotics industry." Bogdanowicz, M. and Desruelle, P. (Eds.), Scientific and Technical Report. 24600 EN, Joint Research Center, Institute for Prospective Technological Studies (IPTS), European Commission.
- Hoffman, H., Pastor, P., Park, D.H., and Schaal, S. 2009. "Biologically-Inspired Dynamical Systems for Movement Generation: Automatic Real-Time Goal Adaptation and Obstacle Avoidance." In *ICRA '09 Conference on Robotics and Automation* (Kobe, May 12-17). IEEE, 2587- 2592.
- Ijspeert, A.J., Nakanishi, J., and Schaal, S. 2002. "Movement Imitation with Nonlinear Dynamical Systems in Humanoid Robots." In *ICRA '09 Conference on Robotics and Automation* (Washington DC, May 12-15). IEEE, 1398-1403.
- Ijspeert, A.J., Nakanishi, J., and Schaal, S. 2003. "Learning Attractor Landscapes for Learning Motor Primitives." In *Advances in Neural Information Processing Systems 15 2003*, Becker, S., Thrun, S., and Obermayer, K (Eds.). Cambridge, MAS: MIT Press, 1547-1554.
- Kalakrishnan, M., Righetti, L., Pastor, P., Schaal, S. 2011. "Learning Force Control Policies for Compliant Manipulation." In *2011 RSJ International Conference on Intelligent Robots and Systems* (San Francisco CA, Sep. 25-30). IEEE, 4639-4644.
- Kappen, H. J. 2007. "An Introduction to Stochastic Control Theory, Path Integrals and Reinforcement Learning." In *Cooperative Behavior in Neural Systems 2007*, Marro, J., Garrido, P. L., and Torres, J. J. (Eds.). *American Institute of Physics Conference Series* 887, 149-181.
- Koenig, N., Mataric, M.M., and Takayama, L. 2010. "Communication and Knowledge Sharing in Human-Robot Interaction and Learning from Demonstration." *Neural Networks* 23, No. 8-9 (Oct-Nov), 1104-1112.
- Kormushev, P., Calinon, S., Caldwell, D.G. 2010. "Robot motor skill coordination with EM-based Reinforcement Learning". In *Intelligent Robots and Systems 2010*, (Taipei, Oct. 18-22). IEEE, 3232-3237.
- Matarić, M.J. 2007. *The Robotics Primer*. Cambridge, Mass: MIT Press.
- Kulvicius, T., Ning, K., Tamosiunaite, M., and Wörgötter, F. 2012. "Joining Movement Sequences: Modified Dynamic Movement Primitives for Robotics Applications Exemplified on Handwriting." *IEEE Transactions on Robotics* 28, No. 1 (Feb), 145-157.
- Nakanishi, J., Ijspeert, A. J., Schaal, S., and Cheng, G. 2004. "Learning Movement Primitives for Imitation Learning in Humanoid Robots." *Journal- Robotics Society of Japan* 22, No. 2, 17-22.
- Pastor, P., Hoffmann, H., Asfour, T., and Schaal, S. 2009. "Learning and Generalization of Motor Skills by Learning from Demonstration." In *International Conference on Robotics and Automation 2009* (Ney York, Jan.1). IEEE, 763-768.
- Pastor, P., Kalarishnan, M., Meier, F., Stulp, F., Buchli, J., Theodorou, E., Schaal, S. 2013. "From Dynamic Movement Primitives to Associative Skill Memories." *Robotics and Autonomous Systems* 6, No. 4 (April), 351-361.
- Sutton, R.S., Barto, A.G. 1998. *Reinforcement Learning: An Introduction*. Cambridge, Mass: MIT Press.
- Tamosiunaite, M., Nemeč, B., Ude, A., and Wörgötter, F. 2011. "Learning to Pour with a Robot Arm Combining Goal and Shape Learning for Dynamic Movement Primitives." *Robotics and Autonomous Systems* 59, No. 11 (Nov), 910-922.
- Theodorou, E., Buchli, J., and Schaal, S. 2010. "Reinforcement Learning of Motor Skills in High Dimensions: a Path Integral Approach." In *ICRA '10 Conference of Robotics and Automation* (Anchorage AK, May 3-7). IEEE, 2397-2403.

AUTHOR BIOGRAPHIES

ANAT HERSHKOVITZ COHEN is a MSc Student in the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva. She received a BSc in Industrial Engineering and Management, also from Ben-Gurion University of the Negev (2013). Her research focuses on robots algorithms that allow robots to learn from humans motions. Her e-mail address: anhe@bgu.ac.il.



SIGAL BERMAN is a senior lecturer in the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva. She received a Ph.D. in Industrial Engineering and Management from the Ben-Gurion University (2002) and a B.Sc. in Electrical and Computer Engineering, The Technion, Haifa. Her research interests include: Human motor control, robotics and telerobotics. Her e-mail address is: sigalbe@bgu.ac.il and her Web-page can be found at <http://www.bgu.ac.il/~sigalbe/>.



Discrete Event Modelling and Simulation in Logistics, Transport and Supply Chain Management

IMPROVING THE DISTRIBUTION PLANNING PROCESS IN THE FOOD&BEVERAGE INDUSTRY: AN EMPIRICAL CASE STUDY

Andrea Bacchetti¹, Massimo Zanardini¹

¹Department of Mechanical and Industrial Engineering, University of Brescia, (BS) Italy,
Brescia, via Branze 38, 25123
andrea.bacchetti@ing.unibs.it
massimo.zanardini@ing.unibs.it

KEYWORDS

Distribution Planning, Case Study, Simulation, Food&Beverage.

ABSTRACT

The distribution planning process is one of the phases of the broader logistics and production planning process for almost every company, and plays a pivotal role in the overall performances (Lee and Kim, 2002; Bard and Nananukul, 2008).

According to Chandra and Fisher (1994), companies can treat this stage in a dual approach. In the first one the overall planning process is considered as an indivisible entity: according to this way several researchers (Glover et al., 1979; Cohen et al., 1988) proposed models in order to coordinate production and distribution activities. In the second approach, the company considers the distribution policy as an independent stage of the entire planning process (for details, see Thomas and Griffin, 1996). Such an approach is more frequently adopted in industry (Chandra and Fisher, 1994).

According to the “independent approach”, this paper illustrates the results of an empirical study involving a relevant food company operating in Italy. The aim of the study is to investigate the distribution planning process, in order to identify the main parameters that govern it, to analyse their impact on the company’s performances and, finally, to propose some improvements, in terms of costs reduction.

According to these objectives, the study addressed, through an intensive case study, two main aspects: (i) the analysis of the company as-is context, encompassing the order process management and the supply chain structure, and (ii) the development of a simulation model that replicates the as-is context and proposes alternative scenarios (to-be), following some *ad-hoc* optimization rules.

Thanks to the simulations, we carried out an optimal configuration for the process parameters, which guarantee, along with the standardization of the order process management, significant economics savings and

increased effectiveness for the overall distribution planning process.

INTRODUCTION

In our paper we provide the description of the results of an intensive case study involving one of the most representative food company operating in Italy (Company, in the remainder of the paper). It is the Italian subsidiary of a German Group, which employs about 9.000 people with 2.000 million € of revenue. The Company manages three different product categories, encompassing dry references (from ingredients of pastry prepared for cakes), cold references (puddings, panna cotta and yogurt) and frozen references (including 19 tastes of frozen pizza). The categories described above detain a different weight for the Company. The dry category includes more than 70% of total references, correspondingly to 55% of total volume streams; the cold products encompass almost 15% of the references (20% of total annual volumes) and the frozen references represent the remaining 15% of the Company’s list (25% of total annual volumes).

The case study deals with the analysis & re-design of the distribution and delivery activities. In particular, the aims of the study are: to investigate the actual distribution planning process, in order to identify the main parameters that govern it, to analyse their impact on the company’s performances and, finally, to propose some improvements, in terms of costs reduction.

In the next section, the problem is outlined and we described the reasons why we adopted a specific analysis methodology. The As-Is Company context, along with the main criticalities, is shown in the third section. In section 4 is depicted the adopted case study methodology; in section 5 are discussed the main elements of the simulation model developed and are reported the simulation results and the achieved benefits. Concluding remarks are drawn in section 6.

PROBLEM STATEMENT AND BOUNDARIES

The scope of the study deals with a specific stage of the planning process, the distribution planning process, one

of the phases in the context of the broader logistics and production planning process, depicted in Figure 1. The company aims to manage in a better way the distribution and delivery process for the frozen pizza, which represents a growing relevant market in the Italian economy.

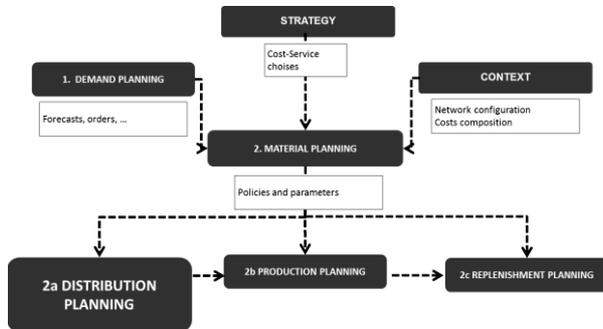


Figure 1: Planning Process Scheme

The integration of production and distribution decisions presents a challenging problem for manufacturers trying to optimize their supply chain (Bard and Nananukul, 2008). The literature provides several research streams for the optimization of the overall planning process. The main findings highlight the differences between an integrated approach, that guarantees a coordinated production and distribution policies, compared to the decoupled approach, where the production and distribution optimization are made separately (Bard and Nananukul, 2008).

Chandra and Fisher (1994) described the benefits and the disadvantages of these two approaches. The first approach deems the overall planning process as an indivisible entity: several researchers (Glover et al., 1979; Cohen et al., 1988; Chien et al., 1989; Thomas and Griffin, 1996) in past years have proposed computational studies to validate models and algorithms in order to coordinate products' production and distribution activities. This approach considers the correlations and the dependencies between production, inventories and distribution stages, ensuring a more coordinated and integrated design and control of these activities, that lead to an increased service level and low cost for the companies (Thomas and Griffin, 1996). Addressing production, inventory and distribution components in a single framework offers an holistic view of the logistics network and provides a good starting point for the full integration of the supply chain (Bard and Nananukul, 2008).

According to the second approach described by Chandra and Fisher (1994), in this study we consider the distribution policy as an independent stage of the entire planning process of the company: we treated company strategy (cost-service choices), company demand planning & forecasting process, production planning and replenishment planning activities as a facts of the context, exogenous to our analysis. Such an approach is more frequently used in industry: first, the company defines the costs policies and determines a production

schedule that minimizes setup and inventory holding costs, and then the distribution policy is developed to satisfy customers' need. In this way, production scheduling and distribution planning are considered as separated problems, which can be solved separately (Chen and Smith, 2010).

There are at least three main reasons why we adopt the independent approach to develop our case study:

1. The production of the frozen products is made in the German plants, that are under the headquarters control. So, the production stage cannot be managed by the Company and therefore could be treated as an independent stage;
2. The decoupled approach works well if there is sufficient finished goods inventory to buffer the production and distribution operations from each other (Chandra and Fisher, 1994). As described in the next section, the inventory level is not a constraint, because the German plants realize the frozen products for all the European country, and the stored references are always available.
3. Finally, the frozen references have a long lifespan and shelf life, reducing the Company's needs to react very quickly to the production phase (Chen and Smith, 2010).

So, we do provide neither solutions nor changes related to the supply chain configuration (number of levels and number of nodes at each level) or stock level: the main goal of the project is to determine which products, how and when will be delivered and transported from the factory (or warehouses) to customers, with the aim to minimize the total cost of the process, ensuring the same service level agreement and respecting the existent constraints.

Obviously, we aware that the results represent just a local optimum for the distribution process, not a global optimum for the overall Company planning process.

THE AS-IS COMPANY CONTEXT

In order to describe in depth the initial situation (As-Is), we analyse two different elements: the supply chain structure and the evaluation order process followed by the logistics employees to fulfil the customers' orders.

The supply chain through which products are delivered in Italy is composed as follows:

1. Two German plants that realize 19 references (different flavours of pizza), stored into two different warehouses even in Germany (closed to the production plants). These sites can be considered the first supply chain level.
2. The second level of the supply chain encompasses two different Italian warehouses, one located in the north and the other in the middle of the peninsula. The company does not own these nodes.
3. At the third level, we find the customers, represented by several (almost 200) stores of the mass retail channel operators.

Due to the supply chain configuration, there are two possible delivery channels:

- Direct way: when a customer order fulfils a series of parameters (see Table 1), the required customers' quantity can be replenished directly from the German warehouse to the retail store. Depending on customer location, it could be replenished both in one day (north clients) and two days (middle and south clients). Trucks starting from German warehouses must be at full load. According to this constraint, we identify two different direct delivery ways:
 - One step direct delivery: when customer requires exactly 33 pallets or multiples (fully loaded), the delivery provides only one unloading of goods at the retail store;
 - Two steps direct delivery: when customer does not require a full load, the Company has to saturate the truck with other references. In this case, in addition to the customer unload, there will be planned an unload to the Italian warehouses for these added references.
- Indirect way: whenever even one of the parameters of the evaluating flow chart is not respected, the order has to be replenished through the Italian warehouses (depending on the customer location). Every customer could be replenished in one day.

The supply chain structure and the delivery ways are depicted in Figure 2.

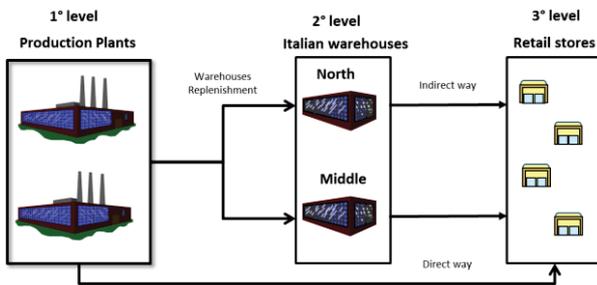


Figure 2: Supply Chain Structure & Delivery Ways

The As-Is evaluation order process, exploited by logistic employees to identify one of the two possible ways to replenish the customers, counts 7 major assessment stages: for every one of these, the logistics operators have to evaluate a specific parameter, and only if it satisfies the threshold value, the evaluation process will continue. At the end of the process (to repeat for each order), if all the 7 checks are positive, the order will be replenished through a direct delivery; otherwise the order will follow the indirect way. In Table 1 are summarized the evaluation phases, with the description of the value parameters, namely the thresholds that permit to discriminate among direct or indirect delivery.

Table 1: Order Evaluation Steps with As-Is Values / Thresholds

Evaluation phase	Parameter	Value / Threshold
1	Order quantity	> 10 pallets
2	Delivery lead time	>= 5 days from the order date
3	Full pallet	<i>Each reference is required in full pallet</i>
4	Geographic localization	< > from islands (Sicily and Sardinia)
5	Available German stock	>= reference order quantity
6	Date of delivery	< > Tuesday
7	Available truck	10 per day: each truck can replenish only one customer (are not allowed cross docking or multi-drop strategies)

In this scenario the Company manifests at least three relevant criticalities:

- the process is not adequately monitored and controlled, and the set of KPIs is quite weak. The Company gathers and monitors only few relevant KPIs (i.e. inventory days in German warehouses), but does not calculate neither the total cost of the process, nor the basic elements cost;
- the evaluating process followed by the Company employees, is not well formalized and standardized (different employee could follow a diverse step sequence to analyse the customer order);
- the values of the parameters that govern the process don't result from empirical analysis, but come from past experience of the Company managers.

OBJECTIVES AND METHODOLOGY

According to the organizational and methodological lacks described above, the main objectives of the project can be summarized as follows:

1. Mapping the order process, drawing all the activities involved in the order fulfilment process, thanks to which standardize the process and identify levers for improvements;
2. Analysis of the supply chain structure, in order to assess which parameters can be considered as constraints or which can be treated as process variables;
3. Designing and developing a simulation model that replicates the As-is scenario, identifying the optimal value of the parameters (described in Table 1) that can be treated and considered as variables endogenous to the Company;

- Evaluating the benefits among the as-is and to-be scenarios, showing to company the time benefits and the economic savings occurred thanks to the changes implemented in the process.

The project follows a case study approach, which allows the investigation of the phenomenon in its natural setting (Voss et al., 2002; Meredith, 1998). Furthermore, the case study methodology allows the questions of why, what and how, to be answered with a relatively full understanding of the nature and complexity of the complete phenomenon.

As case study approach recommend when dealing with complex systems, such as engineering development projects, researchers were “insider” and “participatory” to research (Gosling et al., 2011; Ottoson and Bjork, 2004), to capture depth, nuance, and complex data during the project.

In Figure 3 is drawn the project protocol adopted, described hereafter.

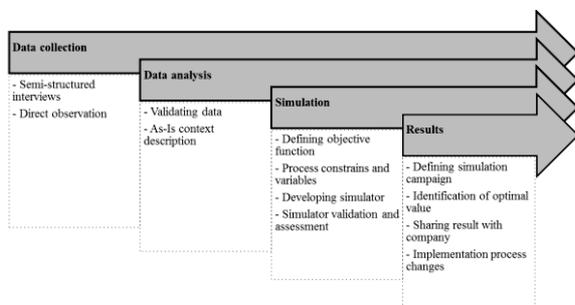


Figure 3: Project Methodology

The first stage consists of gathering data and information about distribution process. Upon the completion of the interviews in which researchers and practitioners collected information involving the Company logistics operators, the researchers analysed and validated the gathered data, assessing the As-is context, from both organizational and methodological standpoints.

Then, the researchers design, develop and validate the process simulations, with the aim of assessing an objective function depending on several parameters and constraints. Company employees and managers were asked to review and validate the simulation model carried out in terms of correctness and completeness. The final version of the simulation model was amended according to the feedbacks received.

As a final stage, a simulation campaign is conducted, in order to investigate the optimal value of the selected parameters that minimize the objective function (described in section 5.1). The evaluation of the intermediate and final results is conducted through several specific workshops with the Company.

PROJECT RESULTS AND MAIN FINDINGS

The simulation model

The simulation model has been developed in ©MS Excel for two main reasons: first of all, the possibility to

exploit the ©Visual Basic for Application programming language in order to reproduce the evaluation processes followed by logistic operator and the supply chain structure; then the general diffusion of the tool and its knowledge by anyone will permit further development and extension of the work by the Company itself.

Discrete event simulation has been used to test the impact of the selected variables and propose their optimal configuration basing on historical demand time series. The simulation model structure is depicted in Figure 4.



Figure 4: Simulation Model Structure

The direct inputs of the simulation are the order series of 2012, encompassing about 6.000 orders and 22.000 order lines, and the supply chain costs structure described hereafter.

The model reproduces the evaluation flow chart presented in section 3 followed by company employees. The fundamental logic can be described as follows:

- The model examines the historical order of 2012 with a daily time bucket, assessing in which of these respect the parameters of the evaluation order process. As a consequence of this assessment process, the simulation model can decide which customers’ orders can be replenished through direct or indirect delivery.
- The model reproduces the material flow towards customers’ retail store, identifying the correct German warehouses where the references (belonging to the considered order) must be taken.
- After having evaluated the daily customers’ orders, the model verifies if there is the need to replenish the Italian warehouses. Such a operation is possible analysing the state of the inventory days related both to the German and Italian warehouses. If there are available trucks and the inventory days in the German warehouses exceed 10, the model calculates the right quantity to send towards the Italian ones.

The objective function is composed of the following cost elements:

- Handling cost:** it manifests when the products are moved out the Italian warehouses, and it is calculated as €/pallet;
- Distribution cost:** it is function of the delivery way assigned to every order. For the direct way it is a fixed unitary cost (€/pallet). According to the first criticality described above, and the general lack of process knowledge and information, the evaluation of this parameter was a significant issue: the Company had only a singular cost element, in which were considered the distribution activities and many other activities related either to different

processes or product category (dry and cold). We disaggregated this cost element until the level of individual activities, and considered the average value as the unitary pallet distribution cost. Indeed, we did not distinguish the cost of delivery to different locations, considering that the customers in the centre and south of Italy are less than 23% and their orders only occasionally meet the requirements for a direct delivery (less than 1% of total deliveries). For the indirect delivery way it depends on the geographic location of the customers and the number of pallets required: a double-entry table is provided by the company;

- **Transfer cost:** this cost deals with the quantity needed to replenish the Italian warehouses from German plants. The method of calculation is the same as the direct way for the distribution cost;
- **Holding cost:** this cost deals with the physical occupation in the Italian warehouses by the company's references (the Italian warehouses are not owned by the company). The cost structure is €/pallet*day.

The output of the simulation campaign consists of a synthetic dashboard, developed in order to implement an appropriate performance measurement system that ensures that actions are aligned to strategies and objectives (Lynch and Cross, 1991; Kennerley and Neely, 2003). The set of KPIs encompasses: the total cost for the company (disaggregated into the four cost elements), the value of the inventory days for the German warehouses and the incidence of direct delivery compared to the total number of delivery.

After having developed the simulation model, we carried out a validation step. A validated model adequately represents the behaviour of the system for the project objectives. Model validation is usually defined to mean "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" (Schlesinger et al., 1979). If a system really exists, the validation step consists of the comparison between the outputs obtained from the simulator with the performance measured in the real system. A model is considered valid for a set of experimental conditions if the model's accuracy is within its acceptable range, which is the amount of accuracy required for the model's intended purpose (Schlesinger et al., 1979).

The total costs obtained running the simulation with all the parameters fixed to the real system values differ from the system performance only about 2%.

Table 2: Comparison Between Real System and Simulation Model

	TOTAL COST	INVENTORY DAYS PLANT 1	INVENTORY DAYS PLANT 2
System performance	€ 3.614.389	10,60	10,00
Simulation result	€ 3.544.452	9,94	10,09
Delta %	-2,0%	-6,2%	+0,9%

Simulation campaign

In order to design the simulation campaign, the evaluation order process parameters have been classified as to which are endogenous and which are exogenous to the company environment. The former are under Company control, so they could be modified and tested during the simulation; the latter are not manipulated by the Company, so in the simulation are treated as constraints.

The parameters classification is presented in Table 3.

Table 3: Process Parameters Classification

Evaluation phase	Parameter	Parameter classification
1	Order quantity	Endogenous
2	Delivery lead time	Headquarter limitation
3	Full pallet	Headquarter limitation
4	Geographic localization	Exogenous
5	Available German stock	Headquarter limitation
6	Date of delivery	Exogenous
7	Available truck	Headquarter limitation

Only one parameter has been classified endogenous: the order quantity respect to which an order can be delivered in direct way from Germany.

The other 6 parameters are not controllable by the Company: 4 of them for specific limitations imposed by the Group Headquarters, while the remaining 2 depend on customers' requirements. Therefore has been tested only the order quantity impact, maintaining others parameters fixed to their real value.

Case study results and insights

The analysis of the 4 cost elements structure, guarantees to identify the hypothetical costs trend moving the order quantity parameter. These evaluations support the empirical decision to carry out a simulation campaign that encompasses 10 simulations: starting from a

minimum of 5 pallets to evaluate a direct delivery, till to 15 pallets (the actual value for the company is 10 pallets). The simulations outputs are summarized in Table 4 and Table 5 and described hereafter.

Table 4 provides the value of the tested parameter (order quantity), the percentage of direct delivery compared to the total shipment, the costs structure for the company, and finally the values of the main constraints for the company, imposed by German HQ to the Company.

All the simulations performed ensure to respect the constraints: the German plants limitations consist of 10 days of maximum inventory days.

Table 4: Simulation Campaign Results

VARIABLE	SHIPMENT COMPOSITION % direct delivery	COMPANY COSTS						MAIN CONSTRAINTS (avg ID < 10,00)	
		Handling Cost	Transfer Cost	Distribution Cost	Holding Cost	Total Cost	Inventory days plant 1	Inventory days plant 2	
Quantity order (pallet) 5	6,4%	€ 159.425	€ 1.729.603	€ 1.242.298	€ 426.842	€ 3.558.167	9,88	10,00	
6	5,9%	€ 156.890	€ 1.702.104	€ 1.259.766	€ 425.691	€ 3.544.451	9,97	9,97	
7	4,9%	€ 157.599	€ 1.709.795	€ 1.255.122	€ 425.764	€ 3.548.280	9,96	10,02	
8	4,4%	€ 157.951	€ 1.713.612	€ 1.252.690	€ 426.115	€ 3.550.367	10,03	10,03	
9	3,9%	€ 158.675	€ 1.721.469	€ 1.248.854	€ 421.307	€ 3.550.306	9,97	9,90	
10	3,7%	€ 159.833	€ 1.734.029	€ 1.240.736	€ 424.624	€ 3.559.222	9,99	9,99	
11	3,2%	€ 161.434	€ 1.751.403	€ 1.225.366	€ 426.076	€ 3.564.279	10,00	10,02	
12	3,0%	€ 161.847	€ 1.755.884	€ 1.223.993	€ 425.032	€ 3.566.757	9,98	9,95	
13	2,7%	€ 163.229	€ 1.770.879	€ 1.216.191	€ 425.100	€ 3.575.400	10,04	10,05	
14	2,5%	€ 163.974	€ 1.778.957	€ 1.209.922	€ 426.219	€ 3.579.071	10,01	10,05	
15	2,3%	€ 164.321	€ 1.782.720	€ 1.205.159	€ 425.978	€ 3.578.178	9,99	9,97	

Table 5: Simulation Campaign Results Compared to Actual Real System (Quantity order = 10) Performances

VARIABLE	SHIPMENT COMPOSITION % direct delivery	COMPANY COSTS						MAIN CONSTRAINTS	
		Handling Cost	Transfer cost	Distribution cost	Holding cost	Total cost	Inventory days plant 1	Inventory days plant 2	
Quantity order (pallet) 5	74,6%	-0,3%	-0,3%	0,1%	0,5%	0,0%	-2,0%	-0,1%	
6	59,6%	-1,8%	-1,8%	1,5%	0,3%	-0,4%	0,2%	-1,0%	
7	34,5%	-1,4%	-1,4%	1,2%	0,3%	-0,3%	0,8%	-1,4%	
8	19,1%	-1,2%	-1,2%	1,0%	0,4%	-0,2%	0,3%	-1,2%	
9	6,9%	-0,7%	-0,7%	0,7%	-0,8%	-0,3%	0,4%	-0,7%	
10	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	
11	-13,2%	1,0%	1,0%	-1,2%	0,3%	0,1%	0,3%	-0,8%	
12	-18,5%	1,3%	1,3%	-1,3%	0,1%	0,2%	-0,4%	-0,7%	
13	-26,4%	2,1%	2,1%	-2,0%	0,1%	0,5%	0,6%	-1,4%	
14	-32,1%	2,6%	2,6%	-2,5%	0,4%	0,6%	0,0%	0,3%	
15	-38,2%	2,8%	2,8%	-2,9%	0,3%	0,5%	0,2%	0,0%	

The behaviour of the total cost trend proves the empirical test range defined before (5-15 pallets) and provides an optimal objective function value in correspondence to the simulation performed with 6 pallets as minimum quantity to evaluate a direct delivery.

We can identify different trends for each cost element:

- **Handling cost:** it is related to the Italian warehouses activities, so when a major number of orders are replenished through the indirect delivery, it increases. Thus, as the order quantity threshold grows, a major number of orders have to be replenished with the indirect mode, so the products are shipped through the Italian warehouses, the ones where this cost manifests.
- **Transfer cost:** decreasing the number of direct deliveries as the order quantity threshold increase, the cost of this element increase because the cost of replenishing the Italian warehouses that have to supply a major number of orders grows.
- **Distribution cost:** this element is subjected to a dual effect. The former is related to the direct deliveries cost: as the order quantity threshold assumes higher values, the number of orders replenished this way decreases. The latter is related to the indirect deliveries cost: as the order quantity threshold grows, the cost of replenishing customers' stores through Italian warehouses increases. The latter effect is less than proportional to the former, so we assist a growth of the total transfer cost moving towards higher values of order quantity threshold.
- **Holding cost:** this element is characterized by non-specific trend. This is reasonable considering that in each simulated scenario the majority of the customers' orders are replenished through the Italian warehouses, and the number of stored pallets doesn't change significantly.

In Figure 5 are shown the cost differences between the optimal scenario (obtained with Quantity Order = 6 pallet) and the simulated As-Is, disaggregated into the 4 cost elements.

Compared to the As-is situation, just changing the order quantity value from 10 to 6 pallets allows Company to reduce the total cost about 2% (approximately 60.000 € per year). This saving, even if it seems low, is actually significant, considering that only one parameter is changed, and the modification does not imply any organizational changes and does not require any ICT integration. It is a real zero-cost modification for the Company.

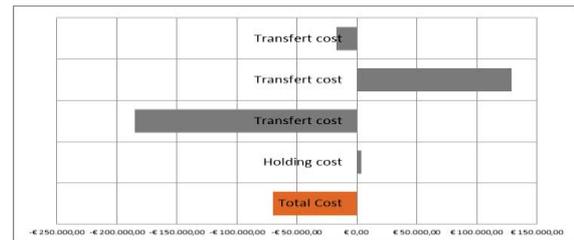


Figure 5: Cost Differences between Optimal Scenario and the As-Is Scenario

Process flow chart reviewing

Another relevant result of the case study has been the reviewing of the flow chart process, with the aim to developing an improved release of the framework, with an increased level of effectiveness. The same number of evaluation steps, but in a different sequence, characterizes the new version: these changes guarantee to the logistic operator to discriminate more quickly the delivery way for each order.

So, in the flow chart, are anticipated the nodes that filter the major number of orders, guaranteeing a more effectiveness evaluation process.

As the result of the simulation campaign proves, the value of order quantity parameter is adapted in the evaluation stage as described in the previous section. The new process sequence is shown in the second column of Table 6.

Table 6: To-Be vs. As-Is Order Evaluation Stages

Order phases As-Is	Order phases To-Be	Parameter
1	1	Order quantity (threshold value: 6 pallets)
2	2	Delivery lead time
3	6	Full pallet
4	3	Geographic localization
5	5	Available German stock
6	4	Date of delivery
7	7	Available truck

We carried out some empirical tests in order to quantify time benefits due to this new configuration of the evaluation order process. The assessment process is composed of three relevant time consuming activities: the first is required for each order, and the remaining two only for the direct deliveries. In order to estimate the economic saving, we considered the modification of incidence of direct delivery compared to the total number of delivery in the As-Is and To-Be scenarios. Table 7 shows that the total evaluation time per order is

about 10,13 minutes in the As-Is situation, while in the To-Be scenario is about 5,59 minutes: with the modification of the order quantity threshold and with the new evaluation order process sequence, the overall Company process achieves an effectiveness 1,44 times higher.

Table 7: To-Be (vs. As-Is) Order Time Evaluation

Activity	Minutes per order (As-Is)	Evaluated order (As-Is)	Minutes per order (To-Be)	Evaluated order (To-Be)
Check input order list	10 min	100,0%	5 min	100,0%
Verify available HQ stock	5 min	1,3%	5 min	5,9%
Manage truck for direct delivery	5 min	1,3%	5 min	5,9%
Total time per order	10,13 min		5,59 min	

Due to this time reduction, and knowing the total number of orders (1.750 in 2012) and the average gross logistic operator cost (about 38 €/hour), it is possible to calculate the (second) economic savings, about 5.000 €/years.

CONCLUSION

Our paper aims to underline how distribution planning is a fundamental task for companies that deliver fast-moving consumer goods (FMCG), and its optimization through the use of simulation can lead to significant economics and time saving for the Company.

The case study underscores that is possible to achieve several benefits also considering only the distribution planning stage, separately from the other planning stages. Particularly when the considered products belong to the FMCG: the main characteristics of these references are short shelf life, long life cycle, low price and low added value. Due to these elements, there are relevant buffers and inventories that permit considering the replenishment, production and distribution stages as independent phases in the planning process. More in detail, the benefits achieved by this “independent approach” are related to:

- the simulations prove the benefits changing the order quantity to evaluate a direct way delivery, from 10 to 6 pallets. This modification implies savings about 60.000 € per year.

- the proposed new release of the flow chart aims at representing an initial model to better manage the required activities and leads to an increased efficiency about 1.44 times (savings of 5.000 € per year).

Further guidelines of improvement are required to analyse whether other constraints limit the company performances. The simulations may underline how constraints works, and explain the possible benefits for the Company and so for the Group whether the constraint value will be replaced to other optimal value. For example the inventory days threshold imposed by the German HQ could influence the total cost of the process. In order to prove to the Company how to use the simulation model to identify any further levers for improvement, we simulate the impact on inventory days constraints: moving from 10 days to 11 days, the total process cost decreases about 25.000 € per year.

A further element to analyse could be the delivery lead time, equal to five days in the As-Is context. Reducing this value from 5 to 4 days, thanks to organizational changes related to the order fulfilment, the Company could increase the number of potential direct deliveries about 1.300 units. It should be analysed the impact of this modification, verifying which cost elements increase and which decrease, assessing the total cost variation.

REFERENCES

- Bard J.F., and Nananukul N. 2008. “The integrated production–inventory–distribution–routing problem.” *Journal of Scheduling* 12 (3): 257-280.
- Chandra P. and L. M. Fisher. 1994. “Coordination of production and distribution planning”. *European Journal of Operational Research* 72 (3): 503-517.
- Chen Z., and Smith R.H. 2010. “Integrated Production and Outbound Distribution scheduling: Review and Extensions.” *Operation Research* 58 (1): 130-148
- Chien T.W., Balakrishnan A. and Wong R.T. 1989. “An integrated inventory allocation and vehicle routing problem”. *Transportation Science* 23/2: 67- 76.
- Cohen M.A. and Lee H.L. 1988. “Strategic Analysis of integrated production-distribution systems: Models and methods”. *Operations Research* 36: 216-228.
- Glover G., Jones G., Kamey D., Klingman D., and Mote J. 1979, “An integrated production, distribution, and inventory planning system”. *Interfaces* 915: 21-35.
- Gosling, J., Hewlett, B., and M. Naim. 2011. “A Framework For Categorising Engineer-to-Order Construction Projects.” Paper presented at the

- 27th Annual ARCOM Conference, Bristol, September 5-7.
- Lee Y.H. and Kim S.H. 2002. "Production-distribution planning in supply chain considering capacity constraints". *Computers & Industrial Engineering* 43: 169-190.
- Lynch R.L. and Cross K.F. 1991. "Measure Up – The Essential Guide to Measuring Business Performance". Mandarin, London.
- Kennerley M. and Neely A. 2003. "Measuring performance in a changing business environment". *International Journal of Operations & Production Management*, 23 (2): 213-229.
- McCutcheon, D., and J. Meredith. 1993. "Conducting Case Study Research in Operations Management." *Journal of Operations Management* 11 (3): 239-56.
- Ottoson, S., and E. Bjork. 2004. "Research on Dynamic Systems: Some Considerations." *Technovation* 24 (11): 863-869.
- Thomas D.S., and Griffin P.M. 1996. "Coordinated supply chain management." *European Journal of Operational Research* 94: 1-15
- Schlesinger, S., R. E. Crosbie, R. E. Gagné, G. S. Innis, C. S. Lalwani, J. Loch, R. J. Sylvester, R. D. Wright, N. Kheir, and D. Bartos. 1979. "Terminology for Model Credibility." *Simulation* 32(3):103-104
- Voss C., Tsiriktsis N., and M. Frohlich. 2002. "Case Research in Operations Management." *International Journal of Operations & Production Management* 22 (2): 195-2

collaborates within the INF-OS initiative (www.asapsmf.org), where he carries out scientific dissemination activities and company transfer projects. He conducts research mainly related to Digital Manufacturing, assessing the impact of new digital technologies (3D printing, Internet of things, ...) in the manufacturing companies.

AUTHOR BIOGRAPHIES



ANDREA BACCHETTI is a post-doc fellow at the University of Brescia (Italy); in 2011 he obtained a doctoral degree in Design and Management of Production and Logistic Systems at the same university with a thesis about spare parts planning. His main research involves supply chain management, demand forecasting, inventory management and services operations management. He is a member of the Supply Chain and Service Management Research Centre (www.scsm.it) at the same University. He is also owner of the INF-OS initiative (www.inf-os.it), where he carries out scientific dissemination activities (e.g. workshop) and company transfer projects concerning the ICT support on business processes.



MASSIMO ZANARDINI is a PhD student at the University of Brescia (Italy). He graduated in March 2012 in Industrial Engineering and is now member of the Supply Chain and Service Management Research Centre (www.scsm.it). He

A TIMED PETRI NET MODEL FOR THE QUAY CRANE SCHEDULING PROBLEM

Roberto Trunfio

LabDoc - Dipartimento LISE
Università della Calabria
Rende (CS), Italy
roberto.trunfio@unical.it

KEYWORDS

Timed Petri Net; Simulation; Maritime Container Terminal; Scheduling; Quay Cranes

ABSTRACT

This paper deals with the problem of constructing the schedule for the operations of a group of quay cranes devoted to discharge/load a set of groups of containers from a vessel at a maritime container terminal. The schedule is constructed starting from the assignment of each individual group of containers to a quay crane under the goal of minimizing the overall vessel completion time, aka the *makespan*. The assignment is provided, e.g., by the search process of an optimization algorithm designed for solving the so called *quay crane scheduling problem*. In this paper, a novel Timed Petri Net model is proposed to construct the schedule from a given assignment. As a novelty, the proposed model considers the initial and final location of the quay cranes to ensure that some necessary physical constraints are satisfied during the idle periods. It also defines an easy-to-implement set of rules to construct the schedule such that the makespan is minimum.

INTRODUCTION

A maritime container terminal (MCT) is a facility located into a port where containers are stored and transhipped between land and ship transports for subsequent transportation. A large number of logistical processes arise in a MCT. In a holistic approach, these processes can be simulated all together by resorting to a high level simulation framework (Legato et al. 2008b), which clearly omits some details from the terminal activities to keep the computational tractability of the model. In fact, currently a detailed representation of all the features of the logistical processes can be obtained by developing a specialized simulation model focused on a single logistical process.

Since the core container transport mode is via ship, the operations focused on the vessels are of primary interest for the terminal managers focused on minimizing the lead time (Steenken et al. 2004). In fact, in literature one of the most studied logistical problems is related to the discharge/loading (D/L) operations of a vessel (Stahlbock and Voß 2008). Vessel D/L operations are performed by resorting to a pool of quay

cranes (QCs) that travel on rails. Each vessel is divided into bays and each bay can be partitioned in two areas located below and above the deck, respectively. Each container is located within a bay and must be discharged or loaded at the port of call according to a pre-defined stowage plan. The problem of indentifying the optimal sequence of the container D/L operations under the objective that the overall vessel operations (i.e. the *makespan*) is minimized; it is known in literature as the *quay crane scheduling problem* (QCSP) (Daganzo 1989; Kim and Park 2004).

When an optimization algorithm is constructed to solve a specific QCSP formulation, a method to evaluate the schedule is required. A schedule is a sequence of activities and events (Pinedo 2002), thus an event-based simulation model can accomplish the task of evaluating the makespan of a schedule for the QCSP. For instance, this approach has been pursued in (Legato et al. 2012; Legato and Trunfio 2013; Trunfio 2014).

Timed Petri Net (TPN) models are an effective and powerful modeling tool to be used in this context (Zurawski 1994; Mejía and Montoya, 2010). In fact, TPN has no modeling limitation that would make the results from the simulation model deviate from the necessary model features. Moreover, in one of the latest formulation of the QCSP proposed by Legato et al. (2012), a TPN has been proposed to calculate the schedule makespan as well as the individual completion time of the specific D/L operations, which has been a plus with respect to all the known methodologies.

Recently, Chen et al. (2014) considered the QCSP by taking into account for additional, but necessary features. Therefore, a novel TPN model that can be used to capture these new features is presented in this paper. Moreover, the proposed model: (i) provides an easy-to-understand set of rules to ensure that the computed makespan is minimum with respect to the given assignment; and (ii) defines an easy-to-read description of the QC operations. The proposed TPN model has been validated by comparing the event-list obtained from other simulation methods from the literature on a set of instances from the literature.

The paper is organized as follows. The next section briefly introduces the necessary background on the QCSP. The subsequent section describes the proposed TPN. The second-last section provides a complete

modeling example. The last section concludes the paper.

THE QUAY CRANE SCHEDULING PROBLEM

The QCSP is a scheduling problem that, according to the modern worldview, is tackled by considering (i) the containers from a bay as a unique indivisible group or (ii) the containers from a bay to be partitioned in groups. The first is known as *QCSP for complete bays* and the latter *QCSP for container groups*. In this paper we dealt with the QCSP for container groups, since it is clearly more challenging and realistic. Furthermore, the former is a special case of the latter.

The modern formulation of the QCSP for container groups (or for simplicity only QCSP from now on) has been proposed by Kim and Park (2004) by enriching a *multiple travelling salesman problem* formulation. Therefore, the Kim and Park formulation is \mathcal{NP} -hard. Several refinements have been proposed in the subsequent years; the major refinements are due to Bierwirth and Meisel (2009), who corrected the representation of some features related to the movement and locations of the QCs during the D/L operations. Moreover, (Chen et al. 2014) has pointed out the attention on the final an initial location of the QCs; the proposed formulation defines the optimal schedule under the assumption that the QC movements are unidirectional (Liu et al. 2006), but removes some modeling features useful for constructing the schedule, e.g. the completion time of the specific tasks.

In the QCSP, a group of containers identifies a task. The set of all the tasks is Ω and $|\Omega| = n$. Each task i has a processing time p_i . A task is located in a bay and the corresponding location is l_i . A task must be assigned to exactly one QC from the set Q , such that $|Q| = m$. The containers related to a bay are divided into groups according to the location of the containers in a bay (below or above the deck), the requested operation (discharge or loading) and the stowage plan (e.g. containers from the same bay area that must be discharged first to speed-up the loading operations of another vessel). Clearly, these facts generate precedence relations for the processing of the tasks: from the same bay, discharge tasks must precede loading tasks; discharge (loading) tasks located above the deck must precede (follow) discharge (loading) tasks located below the deck; other precedence relations may be defined if needed (e.g. due to the stowage plan), but only by ensuring that the previous rules hold true. For a given couple (i, j) of tasks i and j , the set of all the precedence relations is Φ . Figure 1 illustrates an example of a vessel to be handled by using two-QCs.

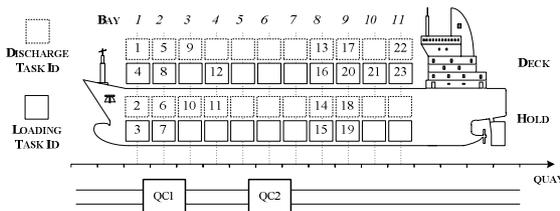


Figure 1: An example of feeder-vessel to D/L with 11 bays, two QCs and 23 tasks.

As shown in Figure 1, the quay side is discretized according to the length of a bay and, consequently, also the travel times of the QCs are calculated according to this space unit. To this extent, let \hat{t} be the travel time required to cover the distance of a bay and $t_{i_l_j}$ the travel time required to move from location l_i to l_j , i.e. $t_{i_l_j} = |l_i - l_j| \cdot \hat{t}$.

The QCs serve along the time multiple vessels from the same berth according to what is defined by the *quay crane assignment and deployment problem* (Legato et al. 2008a; Bierwirth and Meisel 2010). Therefore, with respect to the vessel to be handled, each QC q must perform the assigned tasks within a time window defined by a ready time r_q and a due date d_q . Thus, a QC q has an initial location il_q and a final location fl_q , which are assumed by q at the ready time and when the D/L operations are completed, respectively. Observe that, while the initial location is generally given, the final location is calculated by the QCSP formulation according to the assignment of the tasks to the QCs. For convenience, the start time (i.e. the *zero time*) of the QCSP schedule corresponds to the least ready time of the QCs.

Since the QCs travel on the same rails, then non-crossing constraints must be taken into account. Moreover, to avoid the collision of the QC booms, a safety distance δ must always be guaranteed between adjacent QCs. Clearly, assuming that the m QCs are numbered from 1 to m starting from left-to-right, then two QCs v and w ($v < w$) cannot work simultaneously on two tasks i and j , respectively, if holds true that $l_i > l_j - \delta_{vw}$, where $\delta_{vw} = (\delta + 1) \cdot |v - w|$ (observe that the case for $v > w$ can be easily deduced). From this observation, Bierwirth and Meisel (2009) defined Δ_{ij}^{vw} as the time span to be elapsed between the completion of one between task i and task j and the remaining, under the assumption that i and j are assigned respectively to QCs v and w . Bierwirth and Meisel (2009) shown that only the 4-tuples (i, j, v, w) such that $\Delta_{ij}^{vw} > 0$ have to be considered to define non-crossing and safety distance constraints. Clearly, these constraints must be accounted for in order to guarantee that the schedule is correctly constructed. For convenience, the set of the aforementioned 4-tuples is defined as $\theta = \{(i, j, v, w) \in \Omega^2 \times Q^2 | i < j \wedge \Delta_{ij}^{vw} > 0\}$.

THE TPN MODEL FOR THE QCSP

Petri Nets (PNs) are a powerful and formal modeling language introduced by Petri (1962) which can be represented as a bipartite graph, where nodes are *transitions* or *places* and a transition is connected to a place by an *arc* (and *vice versa*). In simulation, generally a transition represents an event, while a place represents a state. Model changes through the time occurs by using a *token*, a special entity that is put into a place to enable the transitions. A good lecture on modeling with PNs can be read here (Girault and Valk 2003).

There are several interesting extensions of the original language formulated by Petri and TPNs are one of

them. In the literature for the QCSP, a TPN has been proposed in (Legato et al. 2008c) by focusing on the D/L of each single container from a bay and modeling as a plus the operations performed by shuttle vehicles in the quay side; there, however, some important features were not considered (e.g. QC safety distance). Later, another TPN has been used by Legato et al. (2012) to construct the schedule of a given set of task-to-QC assignments. In that TPN model, the deterministic times are associated to transitions and the rules for computing the least cost makespan are regulated by “complicated” equations. We remedy these issues by resorting to a different approach to time modeling. As a matter of fact, another way to introduce time into a PN model is to associate the time with the arcs. Hence, this is the modeling approach pursued in the novel TPN proposed here for the QCSP.

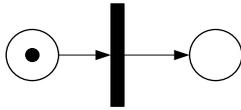


Figure 2: An example of a simple Petri Net: the large circles are places, the vertical black bar is a transition, arcs connect places and transitions and, finally, the small black circle is a token.

We represent the whole process of handling the containers from a vessel as a single-source single-sink TPN. A single token must be put into the source place in order to start the D/L operations. The basic schema of the TPN model is depicted in Figure 3. As shown in the figure, once the schedule evaluation/construction starts (from the “start” transition on), a TPN must be defined for each QC to model a sequence of operations. The QCs have to interact at specific interaction points, hidden from Figure 3, due to the constraints described in the previous section.

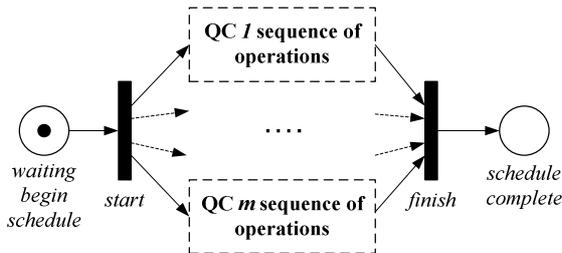


Figure 3: The skeleton of the model.

As suggested by van der Aalst (1996), we represent an *operation* performed by a QC as depicted in Figure 4. As shown in this figure, once that a token leaves the begin place, exactly t unit of times must be elapsed to fire the “finish” transition.

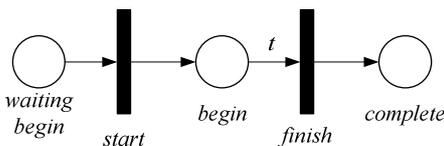


Figure 4: An operation of a QC.

Since the QCs perform the operations in sequence, when two operations are put one before the other in a QC sequence, the “complete” place of the first operation is merged with the “waiting begin” place of the second operation.

We identify four types of operations conducted by a QC during its life cycle: *setup*, *task execution*, *travel* and *completion*. Clearly, for a QC $q \in Q$ there is only one “setup operation” and one “completion operation”, which are put, respectively, at the beginning and at the end of the sequence of operations of q ; in the middle there is the sequence of “task execution operations”, which is defined according to the ordered list S_q of tasks assigned to q . A “travel operation” is put between two “task execution operations”, or a “setup operation” and a “task execution operation”, or a “task execution operation” and a “completion operation”, when needed (i.e. when the travel time is greater than zero). The deterministic time t depicted on the arc that connects the place “begin” with the transition “finish” assumes the value of the ready time, task processing time and travel time for the “setup operation”, a “task execution operation” and a “travel operation”, respectively. The “completion operation” has $t=0$ (therefore it is omitted from the arc). A remark is required for the travel time. When the “setup operation” of QC q and the “task processing operation” for task i are one before the other, then $t = t_{ilql}$; if the “task processing operations” for two tasks i and j are one after the other, then $t = t_{iljl}$; finally, if the “task processing operation” of task i precedes the “complete operation” of QC q , then $t = t_{iflq}$.

Given this premise, we have to model the following features: (i) precedence relations between tasks; (ii) QC non-crossing and safety distance for ready QC; (iii) QC location before the ready period; (iv) QC location after the completion of the last task; (v) QC due date.

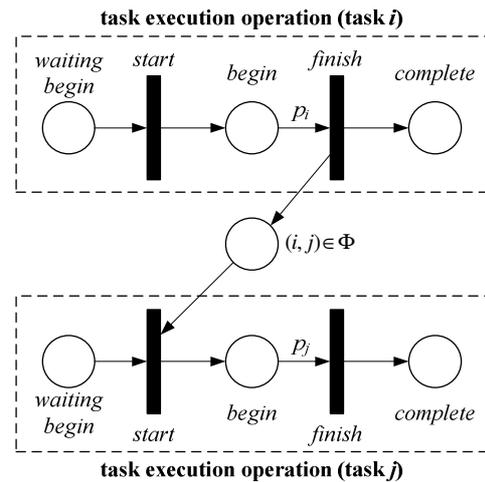


Figure 5: Representation of a precedence relation between two tasks.

Precedence relations between tasks

A precedence relation $(i, j) \in \Phi$, as proposed in (van der Aalst 1996), is modeled as shown in Figure 5. So, to model a precedence relation it is required an ad-

ditional place to be connected between the “finish” transition of the “task execution operation” of task i and the “start” transition of the “task execution operation” of task j , as illustrated in Figure 5. Clearly, if both tasks i and j are assigned to the same QC and i precedes j in the list S_q , then the precedence relations is implicitly accounted for and therefore should not be depicted; otherwise, since j precedes i , the schedule is unfeasible.

QC non-crossing and safety distance

QC non-crossing and safety distance, as explained in the previous section, must be accounted for any 4-tuple $(i, j, v, w) \in \Theta$. Therefore, given two “task execution operations” related to the couple of tasks i and j ($i < j$), which are assigned respectively to QCs v and w and $\Delta_{ij}^{vw} > 0$, the aforementioned constraints can be modeled as provided in Figure 6.

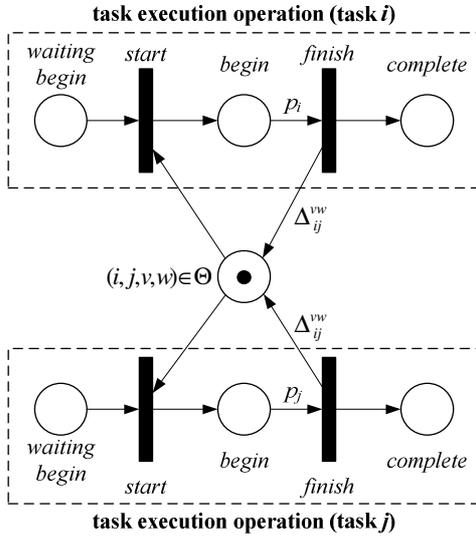


Figure 6: Representation of non-crossing and safety distance constraints.

QC location before the ready period

Since there is no guarantee that the initial location of a QC (i.e. the location assumed from time zero to the ready time) is the same of the location of the first task, then the QC location before the ready period must be accounted in order to avoid QC crossing and to guarantee the safety distance. To model this new feature, we define the set Ψ^I of 3-tuples (i, v, w) such that task i is assigned to QC v and v must wait the ready time of QC w to start working on i . Formally, the new set is defined as $\Psi^I = \{(i, v, w) \in \Omega \times Q^2 \mid i \in S_v \wedge (l_i > il_w - \delta_{vw} \vee l_i < il_w + \delta_{vw})\}$. To ensure that after the ready time of QC w there is no violation of non-crossing and safety distance requirements, the following time span is introduced:

$$\Delta_{ii}^{vw} = \begin{cases} (l_i - il_w + \delta_{vw})\hat{t} & \text{if } v < w \text{ and } l_i > il_w - \delta_{vw}; \\ (il_w - l_i + \delta_{vw})\hat{t} & \text{else if } v > w \text{ and } l_i < il_w + \delta_{vw}; \\ 0 & \text{otherwise.} \end{cases}$$

This stated, Figure 7 illustrates how to model this case for any 3-tuple in set Ψ^I .

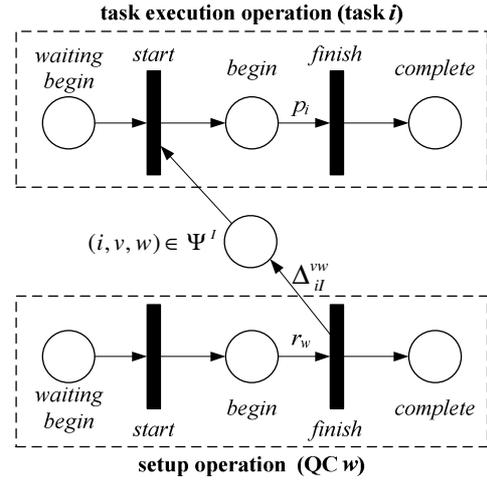


Figure 7: Representation of non-crossing and safety distance constraints for a QC not yet ready.

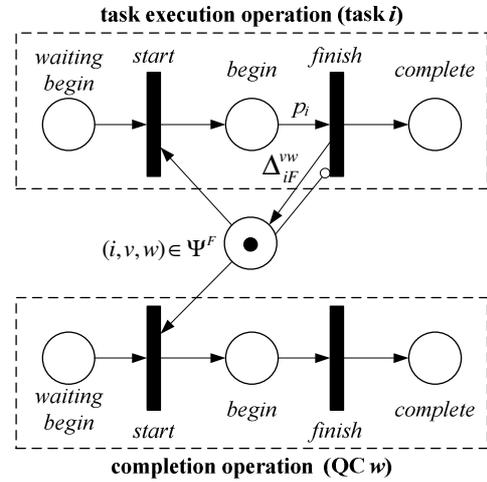


Figure 8: Representation of non-crossing and safety distance constraints for a QC that completed its schedule.

QC location after the completion of the last task

In a similar fashion, also the final location of a QC must be considered to guarantee the non-crossing and safety distance requirements. To this extent, it is introduced a new set, say Ψ^F , that, similarly to the definition of Ψ^I , we define $\Psi^F = \{(i, v, w) \in \Omega \times Q^2 \mid i \in S_v \wedge (l_i > fl_w - \delta_{vw} \vee l_i < fl_w + \delta_{vw})\}$. In this case the time span is defined as follows:

$$\Delta_{ii}^{vw} = \begin{cases} (l_i - fl_w + \delta_{vw})\hat{t} & \text{if } v < w \text{ and } l_i > fl_w - \delta_{vw}; \\ (fl_w - l_i + \delta_{vw})\hat{t} & \text{else if } v > w \text{ and } l_i < fl_w + \delta_{vw}; \\ 0 & \text{otherwise.} \end{cases}$$

We assume that once a QC w completes its operations it becomes definitely idle and cannot be moved along the rails until the completion of the vessel D/L operations. Thus, any assignment of a task i to a QC v such that $(i, v, w) \in \Psi^F$ must be completed before that w repositions to its final location fl_w ; otherwise, the schedule is unfeasible. This new feature disables any “task operation” of a task i assigned to a QC v once that QC w repositions to its final location if and only if

$(i, v, w) \in \Psi^F$. As shown in Figure 8, this new feature is modeled by resorting to an inhibitor arc.

QC due date

For a given QC $q \in Q$, it must be guaranteed that q repositions to the final location fl_q before the due date d_q . So, whenever $d_q < \infty$, we model this requirement as shown in Figure 9. As shown in the figure, after $d_q - r_q$ time units from the ready time of QC q , an inhibitor arc is used to avoid that the schedule is completed. Clearly, from the time d_q the inhibitor arc starts to inhibit the completion of QC q and thus the sink is never reached. As a result, the given schedule is unfeasible.

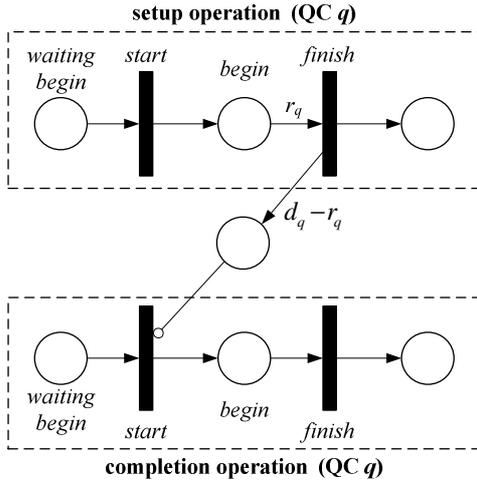


Figure 9: Representation of the due date for a QC.

As one may observe, the previous relation can be established between the “setup operation” and all the following operations: for simplicity, here we depict this requirement only between the “setup operation” and the “completion operation” of each QC.

Time handling and rules for firing

Since the problem at hand requires constructing the schedule such that the makespan is minimum, we introduce some assumptions to define the mechanisms used to handle and to update the simulated time. In particular, we assume that each token in the TPN brings through the graph some necessary information: (i) a variable to memorize the time, say *clock*; and (ii) a variable to save the index of the QC associated to the latest visited QC operation, say *crane*. At time zero, for each token in the TPN is set $clock = 0$ and $crane = null$. A remark for the *clock* variable is due. Every time a token traverses an arc with a weight t , then for this token it is set $clock = clock + t$. Thus, we can observe that each token has its own *clock* value. Whenever a transition fires, a new token is returned for each outgoing arc. For each of these generated tokens, if the traversed transition is from an operation of a QC $q \in Q$, then it is set $crane = q$; otherwise, it is set $crane = null$. Moreover, the *clock* value is set equal to the largest time calculated overall the incoming arcs, where the time is obtained for each arc as the sum of the *clock* values of the token consumed from

the incoming place plus the time from the incoming arc.

An example of this approach is provided in Figure 10. For illustrative purposes only, all the tokens are named (e.g. tok_1, tok_2 , etc). Figure 10 is divided into two parts: the former (Figure 10(a)) and the latter (Figure 10(b)) show the net before and after the firing of the sole transition, respectively. The transition from the figure has three incoming tokens, say tok_1, tok_2 and tok_3 , from three different places. Since the transition is enabled, it may fire, but has to wait until 20 unit of times are elapsed due to the weight of the incoming arc in the middle. Once that the transition fires, two generated tokens, say tok (clearly, they are identical) are put in the outgoing places. The *clock* of each token tok is set such that $clock = \max\{10, 5 + 20, 15\} = 25$. The *crane* variable is set to q , since the transition belongs from an operation of QC q .

Thus stated, the general rule to be applied when multiple transitions are enabled to fire is the “least delay rule”, which implies that, given two transitions, the *clock* value for an outgoing token is temporarily computed, but only the QC with the smaller value of *clock* is selected for firing. If the two *clock* values are equals and there is only one incoming arc for both the transitions, then both the transitions are enabled to fire; otherwise, the transition enabled to fire is selected at random (see for instance the case that models the “QC non-crossing and safety distance”, which implies non-simultaneity between two “task operations” and so, that only one-out-of-two enabled transitions can fire).

Finally, the last firing rule follows here. Given a QC $q \in Q$ such that $d_q < \infty$, if at time d_q there is a token in the “waiting begin” place of the “completion operation”, then the “start” transition of the same operation is enabled to fire; otherwise, the “start” transition becomes inhibited and therefore the schedule becomes unfeasible.

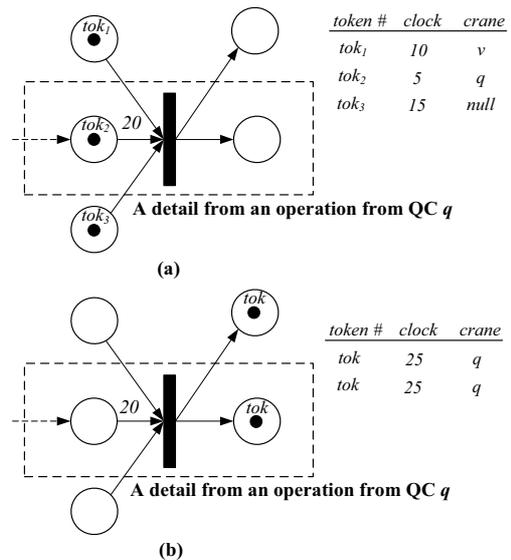


Figure 10: Example of time update. Figure 10(a) and Figure 10(b) show, respectively, the *clock* of the tokens before and after that the transition fires.

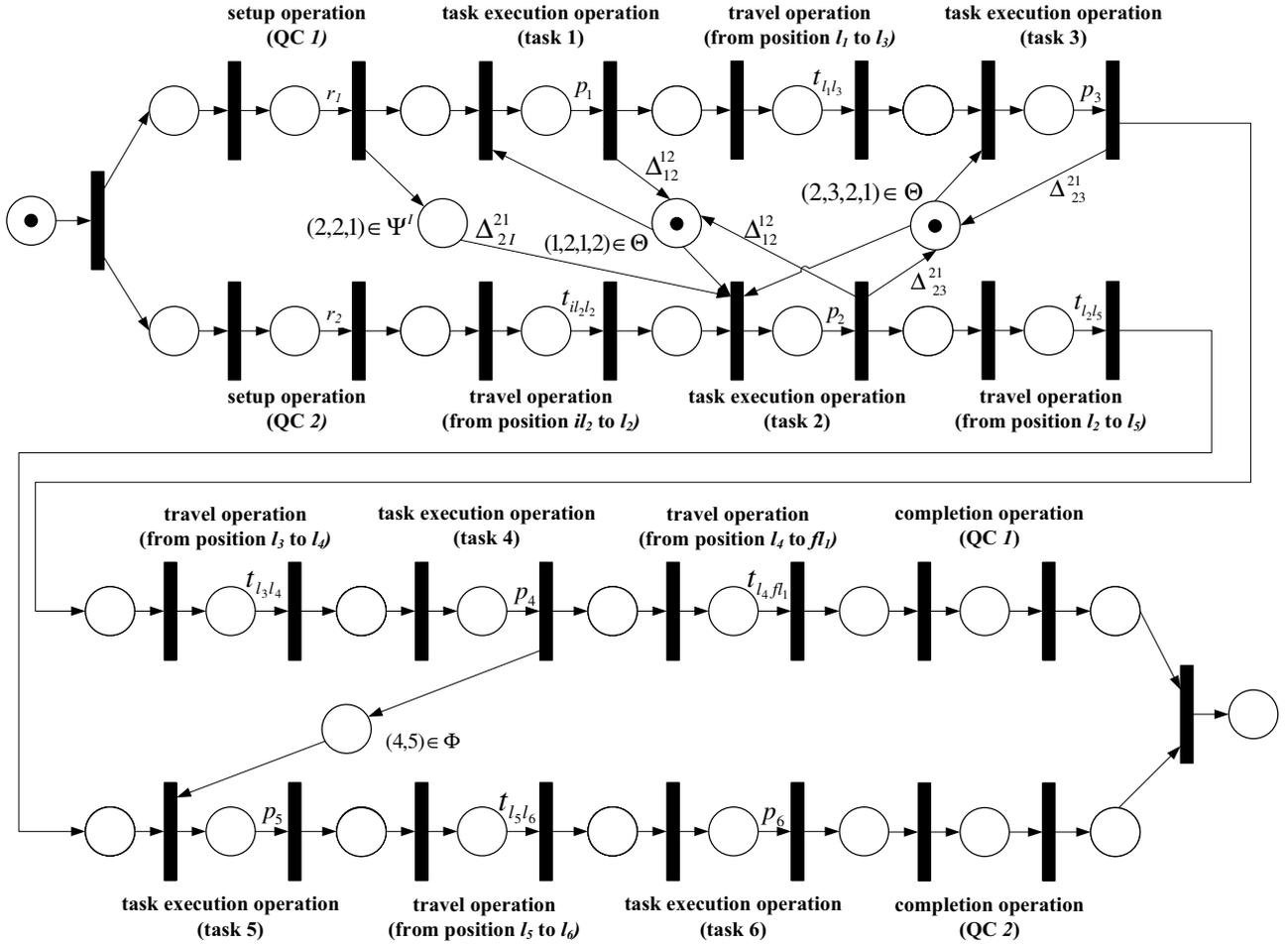


Figure 11: An example of a TPN model for the evaluation of a schedule in the QCSP.

AN EXAMPLE OF THE TPN FOR THE QCSP

An example of a TPN is provided here by starting from the instance illustrated in Figure 8 from Section 3.4 from (Legato et al. 2012). The QCSP instance is defined by the set of tasks $\Omega = \{1, 2, 3, 4, 5, 6\}$, the set of QCs $Q = \{1, 2\}$. The pair of tasks (4, 5) and (5, 6) are, respectively, from the same bay. Furthermore we define the set of precedence relations $\Phi = \{(4, 5), (5, 6)\}$. Moreover, we assume that task locations and the safety margin are given such that $\theta = \{(1, 2, 1, 2), (2, 3, 1, 2), (3, 5, 1, 2), (3, 6, 1, 2), (4, 5, 1, 2), (4, 6, 1, 2)\}$. In addition, we assume that the lists of tasks assigned to the QCs are as follows: $S_1 = \{1, 3, 4\}$ and $S_2 = \{2, 5, 6\}$. The initial location of the QC 1 is in front of the bay of task 1 (which from the 4-tuple $(1, 2, 1, 2) \in \theta$ implies that $(2, 2, 1) \in \Psi^I$, while the initial location of QC 2 is in front of the bay of task 6. For what concerns the final location, we assume that both the QCs comes back to the initial location (i.e. initial and final location coincides): therefore, from $\{(3, 6, 1, 2), (4, 6, 1, 2)\} \in \theta$ (or alternatively from $\{(3, 5, 1, 2), (4, 5, 1, 2)\} \in \theta$) we have that $\{(3, 1, 2), (4, 1, 2)\} \in \Psi^F$. Finally, no due date is set for both the QCs.

Thus, in Figure 11 we report the model that describes this instance. Clearly, processing times as well as travel times are provided as generic values. In the

proposed example, no due date is considered, so the relative modeling features are not reported. Moreover, from S_2 comes that the precedence relation (5, 6) is implicitly accounted and thus the relative constraint is not modeled. Finally, the two 3-tuples in Ψ^F are not considered as well, since, according to the precedence relation (4, 5), both tasks 3 and 4 must be completed by QC 1 before that QC 2 can reposition to its final location.

Observations

Some observations on the proposed TPN model are listed in the following. First, the travel time could be considered in a more concise way, but this would remove some interesting features of the model related to representation of the QC movements along the quay. Moreover, without a doubt, the size of the proposed model can be further reduced, e.g. by removing a lot of couples place-transition that are merely represented to model the *start-finish* events of each operation and are not involved in any kind of constraints (i.e., precedence relations, non-crossing, safety distance). Finally, observe that if the given task lists per QC or the initial/final location of the QCs are not compliant to a well-formed mathematical formulation, then the sink cannot be reached and therefore the corresponding schedule is assumed to be unfeasible. A mathematical

formulation that tackles all these features will be provided in a companion paper.

CONCLUSIONS

We have proposed a methodology for the construction of the schedule for the vessel discharge and loading operations of groups of containers by means of a pool of quay cranes in a maritime container terminal. The methodology is aimed to be used as the makespan evaluation method for the well known *Quay Crane Scheduling Problem*. The methodology consists of a Timed Petri Net model. The proposed model provides a clear representation of the events that occur during the vessel discharge/loading operations. Moreover, it provides a simple set of rules to be used to construct the schedule under the objective of makespan minimization. The proposed TPN model can be implemented and analyzed through any of the several available tools, like *GreatSPN* (Ajmone Marsan 1995; Baarir et al. 2009). Finally, a black-box Java implementation of the simulation model is freely available upon request to the author or at the author's website.

ACKNOWLEDGEMENTS

The author thanks *Prof. Pasquale Legato* (DIMES, Università della Calabria, Italy) for his teaching and mentoring effort that have been a springboard for the writing of this paper. Moreover, the author would like to thank the four anonymous reviewers for their valuable comments and suggestions to improve the paper.

REFERENCES

- S. Baarir, M. Beccuti, D. Cerotti, M. De Pierro, S. Donatelli, G. Franceschinis, "The GreatSPN tool: recent enhancements". *Performance Evaluation Review*, vol. 36 (4), pp. 4–9, 2009.
- C. Bierwirth, F. Meisel, "A fast heuristic for quay crane scheduling with interference constraints". *Journal of Scheduling*, vol. 12, pp. 345–360, 2009.
- C. Bierwirth, F. Meisel, "A survey of berth allocation and quay crane scheduling problems in container terminals". *European Journal of Operational Research*, vol. 202, pp. 615–627, 2010.
- J. H. Chen, D. H. Lee, M. Goh, "An effective mathematical formulation for the unidirectional cluster-based quay crane scheduling problem". *European Journal of Operational Research*, vol. 232(1), pp. 198–208, 2014.
- C. Girault, R. Valk, "Petri Nets for systems engineering: a guide to modeling, verification, and applications". Springer, 2003.
- C. Daganzo, "The crane scheduling problem". *Transportation Research Part B*, vol. 23(3), pp.159–175, 1989.
- K. Kim, Y. Park, "A crane scheduling method for port container terminals". *European Journal of Operations Research*, vol. 156, pp. 752–768, 2004.
- P. Legato, D. Gulli, R. Trunfio, "The quay crane deployment problem at a maritime container terminal". Proc. of the *22th European Conference on Modelling and Simulation* (ECMS 2008), pp 53–59. Nicosia (Cyprus), June 3-6, 2008a. DOI: 10.7148/2008-0053.
- P. Legato, D. Gulli, R. Trunfio, R. Simino, "Simulation at a maritime container terminal: models and computational frameworks". Proc. of the *22th European Conference on Modelling and Simulation* (ECMS 2008), pp 261–269. Nicosia (Cyprus), June 3-6, 2008b. DOI: 10.7148/2008-0261.
- P. Legato, D. Gulli, R. Trunfio, "Modelling, simulation and optimization of logistic systems". Proc. of the *20th European Modeling and Simulation Symposium (Simulation in Industry)* (EMSS 2008), ISBN: 978-88-903724-0-7, pp. 569–578. Amantea (Italy), September 17-19, 2008c.
- P. Legato, R. Trunfio, F. Meisel, "Modeling and solving rich quay crane scheduling problems". *Computers and Operations Research*, vol. 39(9), pp. 2063–2078, 2012. DOI: 10.1016/j.cor.2011.09.025.
- P. Legato, R. Trunfio, "A local branching-based algorithm for the quay crane scheduling problem under unidirectional schedules". *4OR - A Quarterly Journal of Operations Research*, 2013. DOI: 10.1007/s10288-013-0235-2.
- J. Liu, Y.W. Wan, L. Wang, "Quay crane scheduling at container terminals to minimize the maximum relative tardiness of vessel departures". *Naval Research Logistics*, vol. 53(1), pp. 60–74, 2006.
- M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets*. J. Wiley, 1995.
- G. Mejía, C. Montoya, "Applications of resource assignment and scheduling with Petri Nets and heuristic search". *Annals of Operations Research*, vol. 181(1), pp. 795–812, 2010.
- C. A. Petri, "Kommunikation mit automaten". PhD thesis. University of Bonn, 1962.
- M. Pinedo, "Scheduling theory, algorithms, and systems". Prentice Hall, 2002.
- D. Steenken, S. Voß, R. Stahlbock, "Container terminal operation and operations research - a classification and literature review", in *OR Spectrum*, vol. 26, pp. 3-49, 2004.
- R. Stahlbock, S. Voß, "Operations research at container terminals - a literature update", *OR Spectrum*, vol. 30, pp. 1–52, 2008.
- R. Trunfio, "A note on: A modified generalized extremal optimization algorithm for the quay crane scheduling problem with interference constraints", *Engineering Optimization*, *in press*, 2014.
- W. M. P. van der Aalst, "Petri net based scheduling". *OR Spektrum*, vol. 18(4), pp. 219–229, 1996. DOI: 10.1007/BF01540160.
- R. Zurawski, "Petri nets and industrial applications: A tutorial". *IEEE Transactions on Industrial Electronics*, vol. 41(6), pp.567–583, 1994. DOI: 10.1109/41.334574.

AUTHOR BIOGRAPHY



ROBERTO TRUNFIO gained a Ph.D. in Operations Research at the Department of Electronics, Informatics and Systems (DEIS), University of Calabria, Italy, in 2009. He has been senior engineer at NEC where he has worked on optimization and simulation techniques applied to logistics. He currently holds a PostDoc at LabDoc (University of Calabria). His research interests include decision support systems, discrete-event simulation models, simulation-based optimisation, optimization algorithms, semantic web, ontologies, text mining, natural language processing. His home page is at www.roborto.trunfio.it.

DETERMINING TRANSPORTATION MODE CHOICE TO MINIMIZE DISTRIBUTION COST: DIRECT SHIPPING, TRANSIT POINT AND 2-ROUTING

Luca Bertazzi
Department of Economics and Management
University of Brescia
25122, Brescia, Italy
E-mail: bertazzi@eco.unibs.it

Jeffrey Ohlmann
Department of Management Sciences
University of Iowa
IA 52242-1994, Iowa City, USA
E-mail: jeffrey-ohlmann@uiowa.edu

KEYWORDS

logistics, transportation mode, direct shipping, transit point

ABSTRACT

We consider a problem in which a supplier must determine the transportation mode for product deliveries to satisfy demand from a set of retailers. Based on combinations of four possible transportation modes, we consider seven different distribution policies on set of instances derived from data from an Italian company. For three demand scenarios (low, moderate, high), we compare the performance of the various distribution policies. Based on demand characteristics, we characterize the optimal distribution policies. We demonstrate the increase in cost resulting from restricting mode choice to a subset of the possibilities.

INTRODUCTION

Motivated by a situation faced by an Italian company, we consider a problem in which a supplier must determine the transportation mode for product deliveries to a set of retailers. Each retailer's daily demand is known and measured in pallets. The objective is to minimize the overall cost of satisfying retailer demand. For each retailer shipment, the supplier has the choice of four possible transportation modes: *Vehicles*, *Pallets*, *Transit*, *2-Route*. In the *Vehicles* mode, the supplier directly serves a retailer by using a dedicated set of vehicles and pays a fixed cost for each vehicle. In the *Pallets* mode, the supplier directly serves a retailer, but does not reserve the entire vehicle. Instead, the supplier pays a per-pallet transportation cost defined by tiered echelon pricing. The *Transit* mode utilizes a transit point and combines elements of the *Vehicles* and *Pallets* mode. The supplier consolidates the demand intended for a specified set of retailers and uses a dedicated set of vehicles, incurring a fixed cost per vehicle, to deliver this consolidated demand to a transit point. The supplier then pays a per-pallet transportation cost, defined by tiered echelon pricing, to directly deliver each individual retailer's demand from the transit point. Finally, in the *2-Route* mode, the supplier specifies a pair of retailers

and serves them with a route that begins and ends at the supplier, incurring a fixed cost for each vehicle.

In this paper, we compare a variety of distribution policies characterized by the allowed modes. Specifically, we study the following seven policies:

- 1) *Direct Shipping (D)*: Each retailer is directly served using the modes *Vehicles (V)* or *Pallets (P)*;
- 2) *Transit Point (TP)*: Each retailers is served through the transit point using the *Transit* mode;
- 3) *2-Routing (2R)*: Each retailer is served by using either the *Vehicles* or *2-Route* mode (we use a cost structure such that serving a retailer via the *Vehicles* mode is equivalent to serving a retailer via the *2-Route* mode without another paired retailer);
- 4) *D+TP*: Each retailer is served by using the *Vehicles*, *Pallets* or *Transit* mode;
- 5) *D+2R*: Each retailer is served by using the *Vehicles*, *Pallets* or *2-Route* mode;
- 6) *TP+2R*: Each retailer is served by using the *Transit*, *Vehicles* or *2-Route* mode;
- 7) *D+TP+2R*: Each retailer is served by using the *Vehicles*, *Pallets*, *Transit* or *2-Route* mode.

The contribution of this paper is the formulation and application of a series of integer optimization models to compare seven different distribution policies using four possible transportation modes. Specifically, our computational experiments examine the impact of demand level on the distribution policy.

LITERATURE REVIEW

The literature containing elements of transportation mode choice is vast and includes surveys, case studies, and mathematical models. We focus on the most related mathematical treatments of transportation mode choice. For a general survey documenting industry challenges, we refer the reader to Meixell and Norbis (2008).

There is a thread of research which incorporates vehicle routing with mode choice (typically a choice between delivery on routes executed by an internal fleet or delivery via an external carrier). Due to the complexity of the resulting problem (an enriched vehicle routing

problem), this work has focused on the development of heuristic approaches. Côté and Potvin (2009) and Potvin and Naud (2011) are recent works along this meme.

The vehicle routing literature also includes models that incorporate a cross-dock, a type of transit point where loads can be consolidated or separated. Representative work in this area includes Wen et al. (2009) and Tarantilis (2013). These models are concerned with sequencing the loads and the dependencies between the vehicles arriving to and departing from the cross-dock. Due to the complexity of the problem, most studies focus on the development of heuristics.

In contrast to the vehicle routing literature containing elements of mode choice and transit point, our analysis is not saddled with load sequencing decisions. Rather, by our design of the available mode choices, we can focus the impact of demand characteristics on mode selection.

The study of mode choice is also prominent in the inventory literature. Kiesmuller et al. (2005) consider a class of order-up-to policies and compute the optimal policy parameters in the presence of two supply modes. Chiang (2013) considers periodic review inventory policies and computes the optimal policy parameters in the case with two supply modes. In this paper, we are concerned with specifying transportation mode given daily demands from a set of retailers and do not consider the inventory policies of the retailers.

The effect of mode on shipment size has also been examined. Hall (1985) considers a single supplier and single customer to simultaneously determine the optimal mode and shipment size. In related work, Archetti et al. (2011) study lot-sizing in the presence of a transportation cost based on tiered echelons. Our work differs from this thread as we assume demand for our set of suppliers is exogeneous and shipment sizes must satisfy demand.

FORMAL PROBLEM DESCRIPTION

We consider a set $I = \{1, 2, \dots, |I|\}$ of retailers where each retailer $i \in I$ is located at (X_i, Y_i) and has a known daily demand of d_i pallets. The supplier 0 is located at (X_0, Y_0) and a transit point TP is located at (X_{TP}, Y_{TP}) . The supplier can transport the demand of each retailer $i \in I$ via four modes: *Vehicles*, *Pallets*, *Transit* and *2-Route*. The problem is to determine the transportation mode for each retailer $i \in I$ that minimizes the overall transportation cost of satisfying all retailer demand.

In the *Vehicles* mode, a set $J^V = \{1, 2, \dots, |J^V|\}$ of different types of vehicles is available. Each vehicle type

$j \in J^V$ has capacity of r_j^V pallets and incurs a fixed cost c_{ij}^V for each shipment from the supplier to the retailer i .

In the *Pallets* mode, the cost structure is defined by a set $Q^P = \{1, 2, \dots, |Q^P|\}$ of echelons. Each echelon $q \in Q^P$ is defined by a minimum p_q^P and a maximum P_q^P number of pallets and implies a per-pallet transportation cost f_{iq}^P . For $q = 1, 2, \dots, |Q^P| - 1$, $f_{iq}^P > f_{i,q+1}^P$, i.e., the unit cost of each echelon is greater than the unit cost of the next echelon.

In the *Transit* mode, shipment from the supplier to a retailer is broken into two stages. In the first stage, the supplier consolidates shipments for a specified set of retailers and ships the products to the transit point TP on vehicles shared with other retailers. Each vehicle type $j \in J^{TP} = \{1, 2, \dots, |J^{TP}|\}$ has capacity of r_j^{TP} pallets and incurs a fixed cost c_j^{TP} for each journey from the supplier to the transit point. In the second stage of delivery, the shipment from the transit point to each retailer $i \in I$ is based on per-pallet transportation costs defined by a set $Q^{TP} = \{1, 2, \dots, |Q^{TP}|\}$ of echelons. Each echelon $q \in Q^{TP}$ is characterized by a minimum p_q^{TP} and a maximum P_q^{TP} number of pallets and by per-pallet transportation cost f_{iq}^{TP} . For $q = 1, 2, \dots, |Q^{TP}| - 1$, $f_{iq}^{TP} > f_{i,q+1}^{TP}$.

In the *2-Route* mode, the retailer i is served together with another retailer s by using the vehicles $j \in J^{2R} = \{1, 2, \dots, |J^{2R}|\}$ to travel routes starting at the supplier, visiting the two retailers and going back to the supplier. Each vehicle $j \in J^{2R}$ has capacity r_j^{2R} . We denote by c_{isj}^{2R} the cost to serve the retailers i and s together by using one vehicle of type $j \in J^{2R}$.

DIRECT SHIPPING FORMULATION

In the *Direct Shipping* policy, each retailer $i \in I$ is served by using either the *Vehicles* mode or the *Pallets* mode. Let x_i^V be a binary variable equal to 1 if *Vehicles* is used and 0 otherwise, and v_{ij}^D be a non-negative integer variable representing the number of vehicles of type $j \in J^V$ used to serve i . Let x_i^P be a binary variable equal to 1 if *Pallets* is used and 0 otherwise, and z_{iq}^P be a binary variable equal to 1 if echelon q corresponds to the demand d_i and 0 otherwise. Finally, let M be a sufficiently large number. Then, the optimal

Direct Shipping policy can be obtained by solving the following model:

$$\min \sum_{i \in I} \sum_{j \in J^V} c_{ij}^V v_{ij}^V + \sum_{i \in I} \sum_{q \in Q^P} d_i f_{iq}^P z_{iq}^P \quad (1)$$

$$x_i^V + x_i^P = 1 \quad i \in I \quad (2)$$

$$d_i x_i^V \leq \sum_{j \in J^V} r_j^V v_{ij}^V \quad i \in I \quad (3)$$

$$\sum_{q \in Q^P} z_{iq}^P = x_i^P \quad i \in I \quad (4)$$

$$d_i x_i^P \geq p_q^P z_{iq}^P - M(1 - z_{iq}^P) \quad q \in Q^P \quad i \in I \quad (5)$$

$$d_i x_i^P \leq P_q^P z_{iq}^P + M(1 - z_{iq}^P) \quad q \in Q^P \quad i \in I \quad (6)$$

$$v_{ij}^V \geq 0 \text{ integer} \quad j \in J^V \quad i \in I \quad (7)$$

$$z_{iq}^P \in \{0,1\} \quad q \in Q^P \quad i \in I \quad (8)$$

$$x_i^V \in \{0,1\} \quad i \in I \quad (9)$$

$$x_i^P \in \{0,1\} \quad i \in I \quad (10)$$

The objective function (1) expresses the minimization of the sum of the cost of the *Vehicles* and *Pallets* modes. Constraints (2) guarantee that exactly one mode is selected for each retailer i . Constraints (3) guarantee that, in the case *Vehicles* is selected, the total number of vehicles used is sufficient to send the demand d_i . Constraints (4)-(6) manage the *Pallets* mode. In particular, constraints (4) guarantee that at most one echelon is selected. Constraints (5)-(6) specify that the selected echelon q is such that the demand d_i is not lower than the minimum number of pallets and not greater than the maximum number of pallets of the selected echelon. Finally, (7)-(10) define the decision variables of the problem. Note that this problem can be decomposed into $|I|$ problems, one for each retailer $i \in I$.

TRANSIT POINT FORMULATION

In the *Transit Point* policy, first the pallets of several retailers are consolidated and sent to the transit point TP and then they are sent from the transit point to each retailer, separately, on the basis of echelon costs. Let v_j^{TP} be a non-negative integer variable representing the number of vehicles of type $j \in J^{TP}$ used to send the products from the supplier to the transit point and z_{iq}^{TP} be a binary variable equal to 1 if the echelon q corresponds to the demand d_i of retailer $i \in I$ and 0 otherwise. Then, the optimal *Transit Point* policy can be obtained by solving the following model:

$$\min \sum_{j \in J^{TP}} c_j^{TP} v_j^{TP} + \sum_{i \in I} \sum_{q \in Q^{TP}} d_i f_{iq}^{TP} z_{iq}^{TP} \quad (11)$$

$$\sum_{j \in J^{TP}} r_j^{TP} v_j^{TP} \geq \sum_{i \in I} d_i \quad (12)$$

$$\sum_{q \in Q^{TP}} z_{iq}^{TP} = 1 \quad i \in I \quad (13)$$

$$p_q^{TP} z_{iq}^{TP} - M(1 - z_{iq}^{TP}) \leq d_i \quad i \in I \quad q \in Q^{TP} \quad (14)$$

$$P_q^{TP} z_{iq}^{TP} + M(1 - z_{iq}^{TP}) \geq d_i \quad i \in I \quad q \in Q^{TP} \quad (15)$$

$$v_j^{TP} \geq 0 \text{ integer} \quad j \in J^{TP} \quad (16)$$

$$z_{iq}^{TP} \in \{0,1\} \quad i \in I \quad q \in Q^{TP} \quad (17)$$

The objective function (11) expresses the minimization of the sum of the transportation cost from the supplier to the transit point and the transportation cost from the transit point to the retailers. Constraints (12) guarantee that the total number of vehicles used from the supplier to the transit point is sufficient to send the total demand. Constraints (13)-(15) concern the transportation from the transit point to the retailers. In particular, Constraints (13) guarantee that only one echelon is selected for each retailer, while Constraints (14)-(15) guarantee that the echelon selected for each retailer $i \in I$ is such that the demand d_i is not lower than the minimum number of pallets and not greater than the maximum number of pallets of the selected echelon. Finally, Constraints (16)-(17) define the decision variables of the problem.

2-ROUTING FORMULATION

In the *2-Routing* policy, each retailer $i \in I$ is either served directly by using the *Vehicles* mode or together with another customer $s \in I$ on a route starting at the supplier, visiting the two customers and returning to the supplier. We define v_{isj}^{2R} a non-negative integer variable representing the number of vehicles of type $j \in J^{2R}$ used to jointly serve i and s and y_{is}^{2R} a binary variable equal to 1 if the retailers i and s are served together and 0 otherwise. We define δ_{isj}^{2R} equal to $0.5c_{isj}^{2R}$ when $i \neq s$ and equal to $c_{ij}^{2R} = c_{ij}^V$ otherwise. Then, the optimal *2-Routing* policy can be obtained by solving the following model:

$$\min \sum_{i \in I} \sum_{s \in I} \sum_{j \in J^{2R}} \delta_{isj}^{2R} v_{isj}^{2R} \quad (18)$$

$$(d_i + d_s) y_{is}^{2R} \leq \sum_{j \in J^{2R}} r_j^{2R} v_{isj}^{2R} \quad i \in I \quad s \in I, i \neq s \quad (19)$$

$$d_i y_{ii}^{2R} \leq \sum_{j \in J^{2R}} r_j^{2R} v_{ijj}^{2R} \quad i \in I \quad (20)$$

$$\sum_{s \in I} y_{is}^{2R} = 1 \quad i \in I \quad (21)$$

$$y_{is}^{2R} = y_{si}^{2R} \quad i \in I \quad s \in I \quad (22)$$

$$v_{isj}^{2R} \geq 0 \text{ integer} \quad i \in I \quad s \in I \quad j \in J^{2R} \quad (23)$$

$$y_{is}^{2R} \in \{0,1\} \quad i \in I \quad s \in I \quad (24)$$

The objective function (18) expresses the minimization of the transportation cost for serving one or two retailers in every journey. Constraints (19) guarantee that, if the retailers i and s are served together, then the total transportation capacity is sufficient to load the total demand $d_i + d_s$. Constraints (20) guarantee that, if only the retailer i is served on a journey, then the total transportation capacity is sufficient to load the demand d_i . Constraints (21) guarantee that either each retailer is served directly or together with another retailer. Constraints (22) guarantee that if i is served together with s , then s is served together with i . Finally, Constraints (23)-(24) define the decision variables of the problem.

FORMULATIONS OF INTEGRATED POLICIES

We now combine the previous three models to obtain the integrated policies $D+TP$, $D+2R$, $TP+2R$ and $D+TP+2R$. We first formulate the model for the latter policy and then we obtain the others by solving particular cases of this model. Let x_i^{TP} be a binary variable equal to 1 if the retailer $i \in I$ is served by applying the transportation mode *Transit* and 0 otherwise.

Then, the optimal $D+TP+2R$ policy can be obtained by solving the following model:

$$\begin{aligned} \min \quad & \sum_{i \in I} \sum_{j \in J^V} c_{ij}^V v_{ij}^V + \sum_{i \in I} \sum_{q \in Q^P} d_i f_{iq}^P z_{iq}^P + \sum_{j \in J^{TP}} c_j^{TP} v_j^{TP} + \\ & + \sum_{i \in I} \sum_{q \in Q^{TP}} d_i f_{iq}^{TP} z_{iq}^{TP} + \sum_{i \in I} \sum_{s \in I} \sum_{j \in J^{2R}} \delta_{isj}^{2R} v_{isj}^{2R} \end{aligned} \quad (24)$$

$$x_i^V + x_i^P + x_i^{TP} + \sum_{s \in I} y_{is}^{2R} = 1 \quad i \in I \quad (25)$$

(3)-(10)

(14)-(17)

$$\sum_{j \in J^{TP}} r_j^{TP} v_j^{TP} \geq \sum_{i \in I} d_i x_i^{TP} \quad (26)$$

$$\sum_{q \in Q^{TP}} z_{iq}^{TP} = x_i^{TP} \quad i \in I \quad (27)$$

$$x_i^{TP} \in \{0,1\} \quad i \in I \quad (28)$$

(19)-(24)

The objective function (24) is given by the sum of the objective functions of the previous models. Constraints (25) guarantee that for each retailer $i \in I$ exactly one transportation mode is selected. Constraints (3)-(10) are the constraints of the *Vehicles* and *Pallets* modes. Constraints (14)-(17) and (26)-(28) are the constraints of the *Transit* mode. Finally, Constraints (19)-(24) are the

constraints of the *2-Route* mode. We use this model to also obtain the following integrated policies:

1) *Direct + Transit Point (D+TP)* by setting $y_{is}^{2R} = 0$ for all i and s ;

2) *Direct + 2-Routing (D+2R)* by setting $x_i^{TP} = 0$ for all i ;

3) *Transit Point + 2-Routing (TP+2R)* by setting $x_i^V = x_i^P = 0$ for all i .

COMPUTATIONAL RESULTS

We generate realistic problem instances based on data from a primary Italian company. In the *Vehicles* mode, there are two different types of vehicles available at the supplier, i.e., $|J^V|=2$. The first vehicle type has pallet capacity $r_1^V = 20$ while the second has pallet capacity $r_2^V = 34$. The cost structure for the *Vehicles* mode is set so the fixed costs c_{i1}^V is equal to twice the Euclidean distance of customer i from the supplier, while $c_{i2}^V = 1.5c_{i1}^V$.

In the *Pallets* mode, there is a set Q^P of four echelons. Each echelon q is defined by a range $[P_q^P, P_q^P]$ for which the echelon pricing is effective. The four echelon ranges are: $[1,10]$; $[11,15]$; $[16,20]$; $[21, \infty)$. The corresponding per-pallet cost f_{iq}^P is defined such that the cost of one pallet in the fourth echelon is 20% more than $c_{i1}^V / 20$ and the per-pallet cost of each echelon q is 20% more than the per-pallet cost of the echelon $q+1$, for $q=1,2,3$.

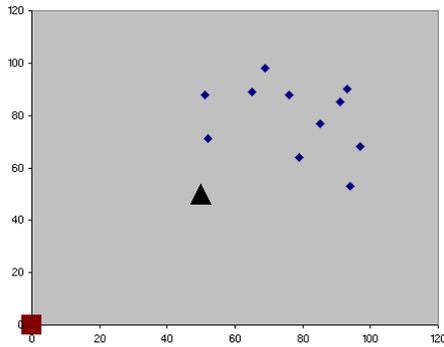
In the *Transit* mode, there are two types of vehicles ($|J^{TP}|=2$) with pallet capacity $r_1^{TP} = 20$ and $r_2^{TP} = 34$, handling the transportation from the supplier to the transit point. The first vehicle type's fixed cost c_1^{TP} is equal to two times the Euclidean distance between the supplier and the transit point. For the second vehicle type, the fixed cost is $c_2^{TP} = 1.5c_1^{TP}$. A set Q^{TP} of four echelons exists for shipments from the transit point to the retailers. Each echelon q is defined by a range $[P_q^{TP}, P_q^{TP}]$ for which the echelon pricing is effective. The four echelon ranges are: $[1,4]$; $[5,10]$; $[11,20]$; $[21, \infty)$. The corresponding per-pallet cost f_{iq}^{TP} is defined such that the cost of one pallet in the fourth echelon is 20% more than $(c_{i1}^V - c_1^{TP}) / 20$ and the per-pallet cost of each echelon q is 20% more than the per-pallet cost of the echelon $q+1$, for $q=1,2,3$.

In the *2-Route* mode, two types of vehicles, $|J^{2R}|=2$, with pallet capacity $r_1^{2R} = 20$ and $r_2^{2R} = 34$. For the

first vehicle type, the cost c_{is1}^{2R} to serve retailers i and s together is equal to the total Euclidean distance of the minimum-length Hamiltonian circuit on the supplier, retailer i , and retailer s . For the second vehicle type, $c_{is2}^{2R} = 1.5c_{is1}^{2R}$.

As shown in Figure 1, the supplier, located in (0,0), serves a set I of 11 retailers, randomly generated in the square having south-west corner in (50,50) and edge of length 50. The transit point is located in (50,50).

Figure 1: Location of the supplier (square), of the transit point (triangle) and of the retailers (rhombuses)



We generate the daily demand d_i for each retailer i as follows. First, a random number $0 \leq \alpha_i \leq 1$ is generated from a uniform distribution. If $\alpha_i \leq 0.5$, then $d_i = 0$; otherwise, d_i is randomly generated from a uniform distribution between a minimum \underline{d} and a maximum \bar{d} number of pallets. We repeat this process for each day of the 24-day horizon so that each day consist of $|I|$ randomly-generate demand values.

We construct demand scenarios by varying the minimum and maximum number of pallets possible:

- 1) *Low demand*: $\underline{d} = 5, \bar{d} = 5$.
- 2) *Moderate demand*: $\underline{d} = 5, \bar{d} = 34$.
- 3) *High demand*: $\underline{d} = 34, \bar{d} = 34$.

For each scenario, we generate five instances and we apply the policies *Direct Shipping (D)*, *Transit Point (TP)*, *2-Routing (2R)*, *D+TP*, *D+2R*, *TP+2R* and *D+TP+2R* by solving the corresponding formulation for each day in the 24-day horizon. We implement the models formulated in the previous sections in MPL (Maximal Software 2013) and apply Gurobi (Gurobi Optimization 2013), a mathematical programming solver, to obtain the corresponding optimal solutions.

Table 1 shows the average percentage error increase in the transportation cost obtained by the policies D , TP , $2R$, $D+TP$, $D+2R$, $TP+2R$ with respect to the policy $D+TP+2R$, in the three scenarios. In the low demand

scenarios the use of the transit point can significantly reduce the transportation cost with respect to direct shipping and 2-routing. In fact, the policies TP , $D+TP$ and $TP+2R$ give the same cost of the policy $D+TP+2R$, while D has an increase in the cost of about 54% and $2R$ of about 71%. In the high demand scenarios, the use of direct shipping and 2-routing give the optimal cost, while the use of the transit point implies an increase in the cost of about 14%. In the case with moderate demand, a right combination of the different transportation modes is needed to have the minimum cost. In fact, the policy D , TP and $2R$ give an increase in the cost of $D+TP+2R$ of about 19%, 5% and 4%, respectively. However, if these policies are integrated, the increase is substantially reduced: It is about 2% for $D+TP$, about 3.5% for $D+2R$ and about 0.16% for $TP+2R$. Note that even the best of these policies is not able to give the cost of $D+TP+2R$. Therefore, the four transportation modes are needed to obtain the best policy.

Table 1. Performance of the Policies with Respect to $D+TP+2R$

Policy\ Scenario	Low Demand	Moderate Demand	High Demand
D	53.71	19.11	0.00
TP	0.00	4.96	14.09
$2R$	70.99	3.88	0.00
$D+TP$	0.00	1.99	0.00
$D+2R$	53.68	3.49	0.00
$TP+2R$	0.00	0.16	0.00

Table 2 illustrates how the different transportation modes (*Vehicles*, *Pallets*, *Transit* and *2-Route*) are used in the optimal solution of $D+TP+2R$. In the case of low demand, all retailers are served via transit point. In the case of high demand, all retailers are served by direct shipping. In the case of moderate demand, all transportation modes are used; about 49% of the retailers are served via transit point, about 29% are served via 2-routing, about 19% are served with dedicated vehicles, and about 3% are served directly with echelon-based costs.

Table 2. Transportation Modes Used in the Optimal Policy $D+TP+2R$

Policy\ Scenario	Low Demand	Moderate Demand	High Demand
<i>Vehicles</i>	0.00	19.43	100.00
<i>Pallets</i>	0.00	2.67	0.00
<i>Transit</i>	100.00	48.57	0.00
<i>2-Route</i>	0.00	29.33	0.00

CONCLUSION

We investigate the problem of determining transportation mode for deliveries from a single supplier

to multiple retailers. We present a series of mathematical formulations to determine the optimal selection of transportation mode from four choices: 1) direct, dedicated delivery with fixed vehicle costs, 2) direct delivery with volume-based costs defined by echelon-based pricing, 3) delivery via a transit point, 4) delivery via routes consisting of pairs of suppliers.

Our computational results suggest that optimal mode choice is highly sensitive to level of demand. At low demand levels, the use of a transit point allowing consolidation is crucial to obtaining a low-cost solution. In contrast, high demand levels facilitate the use of direct shipping and forcing the use of a transit point causes a 14% increase in cost. Moderate demand levels result in the utilization of all four transportation modes, but restricting the modes to only using a transit point or routing pairs of suppliers results in only a 0.16% increase in cost. Future work could explore the utility of these observations in a solution approach for identifying near-optimal distribution strategies in a large network of retailers for which consideration of all possible modes for the entire set of retailers may be intractable.

In this paper, we consider daily mode assignment in which the transportation mode for each retailer is re-optimized each day from the available choices. Depending on the available mode choices, this may result in a retailer's mode changing day-to-day. Future work may consider incorporating the cost of switching a retailer's mode or fixing a retailer's mode over the planning horizon.

REFERENCES

- Archetti, C., Bertazzi, L., & Speranza, M.G. 2011. "Polynomial cases of the economic lot sizing problem with quantity discounts." Technical report n. 358, Department of Quantitative Methods, University of Brescia, Italy (under second revision in *European Journal of Operational Research*).
- Chiang, Chi. 2013. "Optimal replenishment for a periodic review inventory system with two supply modes." *European Journal of Operational Research* 149, 223-244.
- Côté, J.F. & Potvin, J.Y. 2009. "A tabu search heuristic for the vehicle routing problem with private fleet and common carrier." *European Journal of Operational Research* 62, 326-336.
- Gurobi Optimization, Inc. 2013. "Gurobi optimizer reference manual, version 5.6." <http://www.gurobi.com>.
- Hall, R.W. 1985. "Dependence between shipment size and mode in freight transportation." *Transportation Science* 19(4), 436-444.
- Kiesmuller, G.P., De Kok, A.G., and Fransoo, J.C. (2005). , "Transportation mode selection with positive manufacturing lead time." *Transportation Research Part E: Logistics and Transportation Review* 41(6), 511-530.

Maximal Software, Inc. 2013. "MPL manual." <http://www.maximalsoftware.com>.

Meixell, M.J. & Norbis M. 2008. "A review of the transportation mode choice and carrier selection literature." *International Journal of Logistics Management* 19(2), 183-211.

Potvin, J.Y. & Naud, M.-A. 2011. "Tabu search with ejection chains for the vehicle routing problem with private fleet and common carrier." *Journal of the Operational Research Society* 62(2), 326-336.

Tarantilis, C.D. 2013. "Adaptive multi-restart tabu search algorithm for the vehicle routing problem with cross-docking." *Optimization Letters* 7(7), 1583-1596.

Wen, M, Larsen J., Clausen, J., Cordeau, J.F., & Laporte, G. "Vehicle routing with cross-docking." *Journal of the Operational Research Society* 60(12), 1708-1718.

AUTHOR BIOGRAPHIES



LUCA BERTAZZI is Associate Professor of Operations Research at the University of Brescia, Italy. In 1998, he earned his Ph.D. degree in *Computational Methods for Economic and Financial Forecasting and Decision Making* from the University of Bergamo, Italy. His research is focused on mathematical models and exact and heuristic solution methods for decision-making problems in logistics management. His curriculum vita is available at http://www-c.eco.unibs.it/~bertazzi/bertazzi_ing.pdf.



JEFFREY W. OHLMANN is Associate Professor of Management Sciences in the Tippie College of Business at the University of Iowa. In 2003, he earned his Ph.D. degree in *Industrial and Operations Engineering* from the University of Michigan. His research involves mathematical modeling and algorithmic design for decision-making problems in logistics, agriculture, and sports management. Additional information is available at <http://tippie.uiowa.edu/jeffrey-ohlmann>.

Process modelling and simulation for medication-use process

Johan Royer^{1,3}, Michelle Chabrol¹, Jean-Luc Paris²,

¹ LIMOS CNRS UMR 6158, Blaise Pascal University, 63173 Aubière, FRANCE

² Institut Pascal/UBP/IFMA CNRS UMR 6602, Campus des Cézeaux, 63171 Aubière Cedex, FRANCE

³ CHU de Clermont-Ferrand, 58 Rue de Montalembert, 63000 Clermont-Ferrand, France

E-mail : jroyer@chu-clermontferrand.fr, chabrol@isima.fr, paris@ifma.fr

KEYWORDS

LT, Simulation, Modelling methodology, Medication-use process, BPMN

ABSTRACT

University Hospital of Clermont-Ferrand requested decision-making software in order to help the practitioners to choose a strategy and to make choices. First we have to understand and identify the system with accuracy and choose the best tools to develop this software. Therefore, we need to select a language (BPMN) and a methodology (ASDI) able to model the processes of the system following choices and requests of the hospital. In addition to that, we have to choose the tool (SIMIO) to build our simulation model which is the main part of our decision-making software. To illustrate the result of this work, a quick example of application of this tool is presented below.

INTRODUCTION

Because of the increasing number of constraints – especially economic – decision-makers have a lot of difficulties to change their organization, anticipate the future and decide which strategy will be the most efficient. In this context, unfavorable for investment, University Hospital of Clermont-Ferrand, like all French hospitals, has to confront these problems and thus requested a tool able to help them face these ones.

Indeed, when we focus on medication-use process (MUP), we see that all practitioners who have to make strategic or economic decisions to improve their department's efficiency are in a difficult situation which requires having as much information as possible.

The question addressed here is to carry out a methodology and a tool able to give a lot of information on different domains (economic, human resources, capacity of production, etc.) on a specific field, the medication-use process (from the prescription to the administration).

Moreover, University Hospital of Clermont-Ferrand has still a lot of progress to do in its management and production to significantly improve its efficiency. This is another motivation to have a decision-making tool which enables to test and explore new solutions. For example, among this improvement progress, it is possible to list dispensing robot, individual nominative dispensing, drug prescription validation, etc.

So, the following article presents how we arrive at a decision-making software adapted to pharmacists. This article is organized as follows: the first section describes the context of our work and is followed by a section regarding the tools that we use to model the MUP. Before presenting a case of application of our simulation model in the last section, we present this model and the software used to build it. Finally, we conclude on the current situation of our work.

HOSPITAL REQUESTS

In order to have useful decision-making software, the hospital imposes some conditions and options that the software must have. These requests can be classified in two categories: technical and ergonomic.

Technical requests

For the hospital some options must be included in the software, in order to be the most efficient to test new solutions:

- Production for external medical organization: The tool must be able to support the integration of external structures production. Indeed, nowadays a hospital able to supply medications to other medical organizations has a big economical advantage. The only limit imposed by the hospital (with the production capacity) is to avoid a failure in the cold chain. So, the other structures must be at most one hour of transportation away. So for this request we have two main parameters: distance to hospital (in minutes) and number of demands.
- Type of dispensing: Today, in a hospital, it is possible to find different ways to dispense medicines to patients. Indeed, depending on technological solutions, a hospital can use a robot to prepare the patient prescription while in another one this work is done by pharmaceutical assistants. Below, we give a short description for each type of dispensing:
 - Individual Nominative Dispensing (IND): computerized prescription and dispensing robot;
 - Global Dispensing: handwritten prescription, shelves and bulk delivery to departments.
 - Computerized Dispensing: computerized prescription, shelves and bulk delivery to departments.

These three types are the mains but there are some specific situations such as:

- Automated dispensing cabinets : computerized prescription and unit dose drug dispensing;
- Manual Individual Nominative Dispensing: Contrary to the others, in this situation even if the prescription is computerized, its preparation is entirely manual.

Therefore, we must take in consideration the way to prepare a prescription/order and more precisely how each medicine is supplied.

- **Human resources:** Of course, one of the requests is the capacity to play on number of pharmaceutical assistants, storekeepers, pharmacists or on their versatility. Therefore we have here the following parameters:
 - The number of workers for each profession;
 - The versatility of each worker.
- **Automation:** As said previously, the hospital wants to be able to include robots in its production. So, it is necessary to give them the possibility to define the parameters of robots in the decision-making tools. When we look at dispensing robots, we see that the main characteristics are the capacity of production expressed in doses/hour and the stock capacity (in dose). For unit dose packaging robot, which is also an important tool, the main characteristic is, as the dispensing robot, the capacity of production (in doses/hour).
- **Financial cost:** The last request of the hospital is to know the economic impact of the tested solutions. It is not a parameter needed for simulation but we have to integrate this in the decision-making tool in order to give the most accurate information to decision-makers.

Ergonomic requests

As the tool will be intended to practitioners, namely people with no high skills in informatics, the software must be the simplest to use and the most understandable. Therefore we have here three important criterions:

- Ergonomic interface to enter important data;
- 3D simulation to facilitate the understanding;
- A simple name for each required data.

With these requests, it is now possible to begin the second step of our work in order to build a simulation model: the understanding of the MUP. However, to be able to do this, we have to identify the flows in the system and the involved processes. Indeed, with this work, we will know who are the stakeholders in the MUP, how the activities, the tasks are linked, etc. Finally, we have to build a map of the MUP processes.

HOW TO UNDERSTAND THE MUP ?

In order to study efficiently the MUP, we have to model the processes involved. Nowadays, a lot of languages, methods and methodologies give us the possibility to build such a map. Naturally, each of them has advantages and drawbacks and we must choose one of them.

Even if we can choose a language corresponding to our expectations, we wanted to give to pharmacists the opportunity to help in this work. In addition to that, after some reflections we decide to give to pharmacists not only a decision-making tool, based on a simulation model, but also a background which enables us to have a better comprehension of the MUP possibilities. At this level, doing that implies to explore all MUP configurations and consequently to study many hospitals. Finally, a processes map of University Hospital of Clermont-Ferrand becomes a generic processes map of hospitals, where the maximum of possible configurations for the MUP will be included.

As previously stated, to be able to develop our processes map requires to use a language usable for the pharmacists and for us. To solve this problem we selected five languages: Event-Process Chain, UML, Petri Nets, BPMN and LAESH (Rodier 2010). Each of this language was evaluated by pharmacists and junior pharmacists in order to select the most understandable for them. The result of this evaluation (which is not presented in this article) classed these five languages as follow: 1 – BPMN; 2 – UML; 3 – LAESH; 4 – EPC; 5 – Petri Nets. Therefore we choose to build our processes map with BPMN.

BPMN Model

Using BPMN language enables us to build a model of the MUP from the prescription of medicines to their administration. Therefore, this model incorporates all of the entities interacting with the MUP and maps all involved activities.

It has to be noted that the whole BPMN model is not built with the same granularity (microscopic, mesoscopic, macroscopic). Indeed, some activities are not very important for us and outside of our direct problem. In this case, these activities have a macroscopic representation. Otherwise, the activities that are very important for us and for the understanding of the MUP have a microscopic view.

This BPMN model has been validated by pharmacists.

Figure 1 shows a macroscopic view of pharmacy. This symbolism does not come from BPMN but we developed it in order to simplify the whole map and to understand the main phases in one look.

On Figure 2, a part of the microscopic view of “Phase 1” is presented.

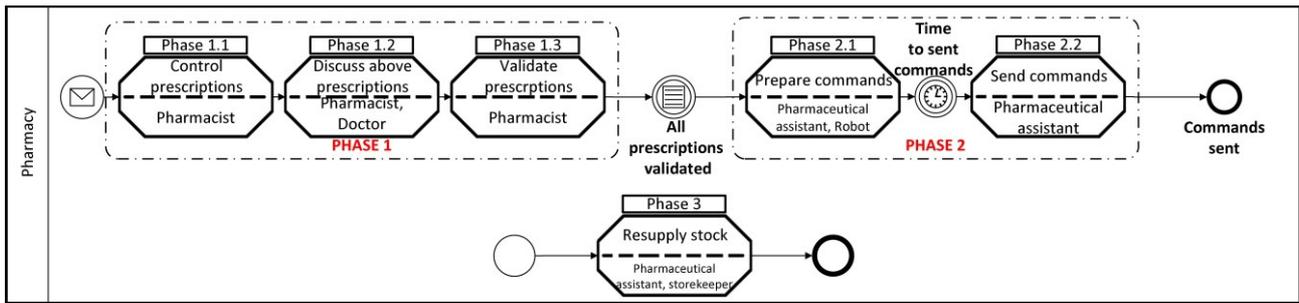


Figure 1 : Macroscopic view of pharmacy

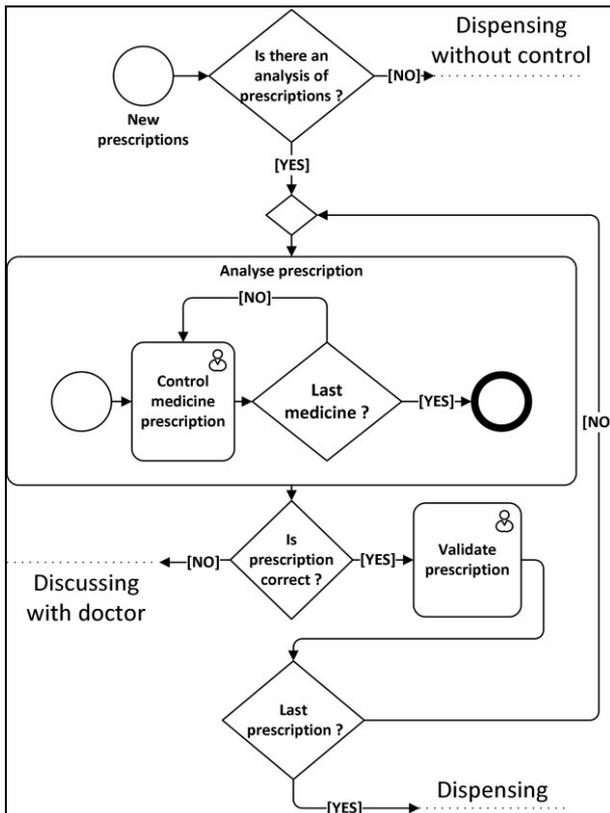


Figure 2 : Part of a microscopic view of "Phase 1"

In our model it is possible to find all actors (human or not) involved in the MUP. These actors are classified in three pools:

- **Department:** Doctor, nurse, pharmacist and dispensing cabinet;
- **Pharmacy software;**
- **Pharmacy:** Pharmacist, Pharmaceutical assistant, storekeeper, robot, mechanical carousel.

As it is impossible to describe the whole model in this article, we will only present a "block description" of this model in order to understand the MUP. Indeed, this model needed 3 months to be developed, including discussion and processes modelling. Furthermore, this processes model has around 150 activities only for standard process and integrate 5 main scenarios (depending on dispensing politic: individual nominative

dispensing, etc.) and a total of 132 different scenarios with the various possibilities.

So in this BPMN model we can find for each pool:

- **Department:** Ordering the prescription and medication reconciliation; Transcribing onto care plan; Checking needed medicines; Ordering needed medicines; Controlling received medicines; Filling and updating pillbox; Administering medicines and treatment updates; Validating the treatment;
- **Pharmacy software:** Entering prescriptions into computer system; Updating robot production plan; Controlling automated dispensing cabinets;
- **Pharmacy:** Reviewing prescriptions; Resupplying stocks; Dispensing medicines; Distributing medicines.

ASDI Methodology

Although some languages, as BPMN, are able to specify and analyze the functioning of a complex system, they are not necessarily included in a method or methodology. Conversely, some methods and methodologies are not necessarily supplemented with a language.

Indeed, a modelling methodology provides a general guide and can be defined as follows:

"A modelling methodology is a set of methods, tools, approaches and concepts for modelling a system" (Chauvet 2009).

Therefore its role is to combine approaches, languages or methods from different backgrounds in order to integrate them in a logical approach.

This is the reason why we chose ASDI methodology. ASDI stands for Analysis, Specification, Design, and Implementation and it is a methodology developed to enable the design of tools for decision support. In this methodology, these tools are called action models (simulation model, metaheuristic model, analytical model, etc.) (Gourgand et al. 1991; Gourgand et al. 1992). This methodology enables to model a class of systems and produces a "generic knowledge model" of this class. A library of software components is built and is used to generate an action model (equivalent to computer program) (Fig. 3). The conceptual framework of the ASDI methodology is based on a modelling

process separating explicitly the collection of knowledge and its utilization. So, each step requires choosing the most appropriate tools based on system characteristics, objectives, etc.

Finally, systemic view, reflecting ASDI application, consists of a hierarchical decomposition of the system with the object paradigm in three interrelated sub-systems (communicating and complementary). They are logical, physical and decision-making subsystems.

We will stop here regarding methodologies, because ASDI was already used successfully in hospitals (Huet 2011; Rodier 2010) owing to its flexibility and adaptability.

But contrary to (Huet 2011), who focused on final decision-making tools, and (Rodier 2010), who used LAESH language and did not work on the MUP, our work gives to practitioners, and especially pharmacists,

an approach to work by themselves with a more common language (BPMN) on the MUP.

With the Figure 3 of ASDI methodology, it is possible to see its main steps. We see that some tools are needed to use ASDI and consequently also for our approach. The first of all is the UML class diagram useful to identify the entities in the system and the links between them. The work of (Huet et al. 2010) and (Rodier 2010) have made possible the construction of our data collection, reusing and completing UML class diagrams previously established.

The second was in original ASDI, a UML activity diagram. Thus, as BPMN is more efficient than UML for our work, we need to replace activity diagram by BPMN model.

As seen previously, the choice of the BPMN language went through an evaluation of several different languages and integrated pharmacists in the development of this first tool.

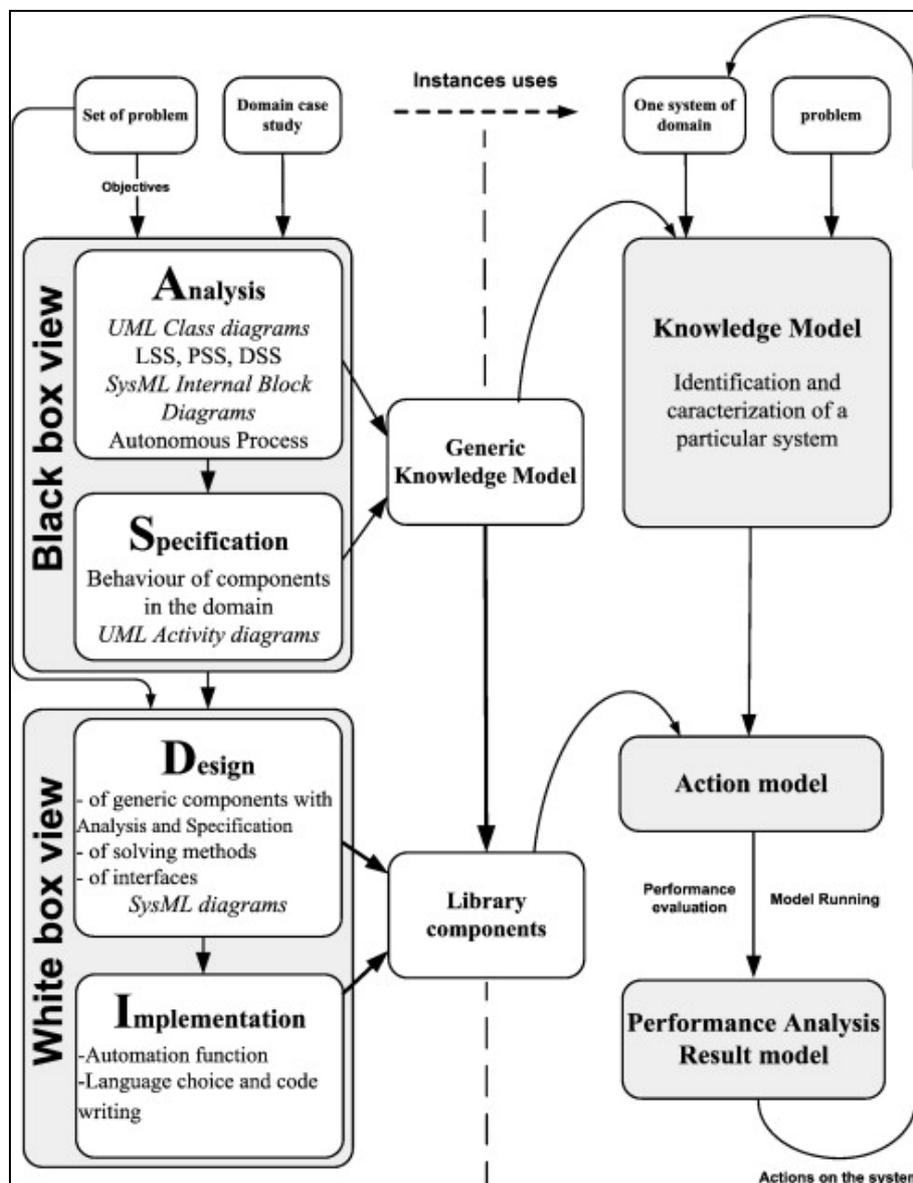


Figure 3 : ASDI methodology

Therefore, their involvement has made possible the development of a strong understandable model to them. Finally, the use of SIMIO software helped to design a generic simulation model of the MUP that can support any different hospital organization; we will now present those.

SIMIO AND SIMULATION MODEL

The choice of software for performing the simulation model is SIMIO.

This software is a tool to build and run dynamic 3D animated models of a wide range of systems – e.g. factories, supply chains, emergency departments, airports, and service systems. To model a system we model the flow of entities through the system and the resources that constrain that flow. SIMIO uses an object approach to modeling, whereby models are built by combining objects that represent the physical components of the systems such as workstations, conveyors, and forklift trucks in a manufacturing facility, or the gurney in an emergency room of a hospital system. Each object has its own custom behavior as defined by its internal model that responds to events in the system.

Moreover, its graphic interface for constructing 3D models is also an asset, the objective being that this model will be used by practitioners.

The model

During the development of our model (Figure 4), this last changed constantly, and so far, it is not yet the final version. Indeed, each discussion with pharmacist gives new ideas to improve it and as hospital world is in constant evolution, it is difficult to not adapt our model to the reality. So, the following presented model is the most complete for the moment but certainly not the last.

What does our model ?

For the time being, the simulation model – stochastic (number of orders, preparation time, number of products, etc.) – is able to simulate the whole preparation of orders (commands, prescriptions, demands) coming from hospitals departments. So, all the steps from prescription to administration can be found in this model.

We also tried to consider all hospitals configurations in the simulation model, but as we have the constraint to be generic, we cannot consider spatial data (at least not yet).

Before giving an example, we will describe all elements included in this model. Each element is identified on Figure 4 with a number.

Entities and human actors

As said previously, we chose to considerate only orders and not the medicines. Indeed, medicines number is always in the thousands, although it depends on the hospital size. So the orders are our first entity. The second is the medicines coming back from departments.

As for medicines in orders we consider a “return list” (No. 2) instead of each product.

At first sight, these choices could seem inappropriate because they do not take the large number of medicine types in consideration. Indeed, some are stored in cold room, while others need special preparation. So to avoid this problem there is a probability for each order to be composed of different types of products. These types are: Manual picking; Automatic picking (dispensing robot); Cold room; Solute/Solution; Carousel picking. So these main types correspond to the different ways to store medicines in pharmacy. Same goes for returned medicines.

In addition to these two entities, human actors are incorporated in our model. Indeed for each activity which needs to be completed by a human, we associate a worker classified in 2 categories: Validator (No. 4) (Junior or Senior Pharmacist); Pharmaceutical assistant (No. 5) (PA).

While validator is needed to control and validate prescription (equivalent to nominative order), pharmaceutical assistant prepares orders with the help of resources that we shall now present.

Resources

The human actors need some resources to work and prepare the orders. Depending on the type of the order, these needs will not be the same. Therefore we have:

- Control area (No.3): where PA edits, controls and validates the orders.
- Validation area (No. 6): where pharmacists work;
- Dispensing robot (No. 7);
- Mechanical carousel (No. 8);
- Shelves: For manual picking (No. 9);
- Cold storage room (No. 10);
- Restock area (No. 11): Fictional stock for returned products.

Processes

The “Processes” are a specificity of SIMIO and are not directly visible for common users. These objects are a sequence of steps that may changes the state of one or more elements in the model to perform some custom logic. This logic can concern seize/release resources, evaluate alternatives, change behavior, etc.

So, we find in our processes:

- Seize/release PA;
- Manage the number of available PA;
- Mobilize PA if needed (with a defined maximum);
- Guide orders in the model;
- Seize/Release/Create Validator;
- Etc.

Options recap

Finally, if we resume all possibilities available in the simulation model, the user can choose:

- The number of each human actor with a schedule for each day (seven slots a day);
- The number of each resource with schedule for some of them (for example dispensing robot);
- The type of each order with parameters (preparation mode, number of product, etc.);
- The global number of orders on the day (12 slots);
- A specific PA for an order;
- Etc.

Data

Collecting data

All data previously needed to run the model can be found on real system and in the Pharma software, used in University Hospital of Clermont-Ferrand. This software was used to retrieve orders delivered during May, 2013.

Representing data

The obtained data were mostly in spreadsheets. So, it was necessary to handle these ones to obtain mathematical equations and distributions to model data like arrival time, number of elements, preparation time, etc. When all data were ready, we build new tables which can be imported in SIMIO. For example, we have tables with number of orders per hour, weekly schedule for PA, type and characteristics of orders, departments delivered per day, etc.

Example of process

Before presenting an application for our simulation model, here is an example for a computerized dispensing, meaning computerized order from department and manual preparation. In addition to that, we suppose that this order must be prepared by the PA #1 and it includes solutions, cold and normal medicines. Therefore, the process of this order (black arrows in Figure 4) is in our model:

- Order is generated in Department (No. 1);
- As order does not include IND, it goes directly in No. 3;
- If the PA #1 is available the “Edit and control” step begins. Otherwise, the order waits for this specific PA;

- After that, as order includes cold medicines and solutions, this order is split in three “sub-orders” (**A** in Figure 4). One goes to No. 10, and as the two others have manual preparation, they go to No. 9;
- At this level, there is no priority between each “sub-order”. So they are prepared one after another exclusively by PA #1 who is dedicated temporarily to these orders;
- When one “sub-order” is ready, it waits in **B** until others are ready, then the model combines and sends them to the department.

CASE OF APPLICATION

To illustrate the application of our simulation model, we test here a simple situation where the University Hospital of Clermont-Ferrand wants to integrate in its pharmacy a dispensing robot in order to prepare prescriptions. We will focus here only on one day (Monday 13th May 2013) for two motives. The first one is that the influence of the robot on the production is more significant on a busy day, where a lot of orders is placed. The second one is that a robot can improve production on a long period, but not equally on each day. Therefore a hospital will not invest in a robot only for some days of improvement. So, to show the real impact of the robot, we need to study each day of production and in our case a day where a robot could be really useful.

During this day, the pharmacy prepared 130 orders for 60 departments in 9 hours of work (including breaks). For now, 6 departments use computerized prescriptions, so if we complete this with a dispensing robot we are in a situation of individual nominative dispensing. With this new organization on Monday 13th May, the number of orders would decrease of 20 orders (110 left). These orders are replaced by 50 prescriptions for 50 patients among all patients of the 6 computerized departments. The Table 1 resumes the situation:

Table 1 : Number of orders/prescriptions to prepare

	Before IND	With IND
Manually	130	110
With robot	0	50

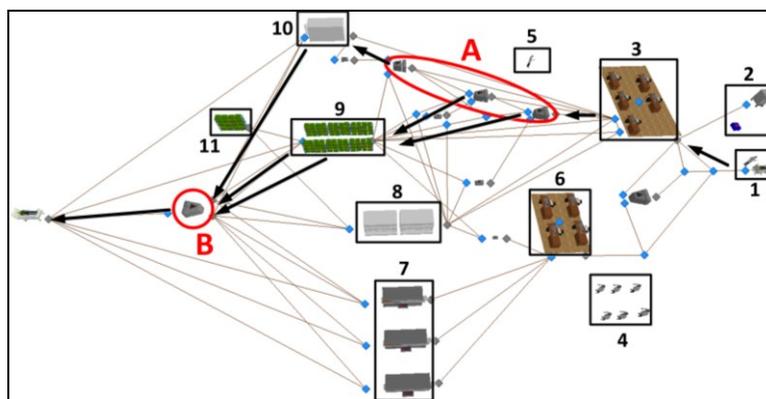


Figure 4 : Graphic view of our simulation model

In addition to that, the robot requires a PA during the peak of production (between 8h and 12h). So the Table 2 gives the number of PA per period on this day:

Table 2 : PA Monday schedule

Period	8h/12h	12h/13h	13h/16h	16h/17h
Before IND	6	1	6	1
With IND	5 + 1 (robot)	1	6	1

We do not focus here on the robot capacity production. Indeed, as we have not a lot of prescriptions to prepare, any robot would be suitable for us. Therefore, when we run our simulation model we obtain the following results:

Table 3 : Number of prepared orders/prescriptions

	At 11h30	At 13h	At 16h	At 17h
Reality	46	80	125	130
Simulation before IND	58	71	116	123
Simulation with IND	51	60	103	109

With the diminution of orders to be prepared with the IND we attain the objective of 110 orders for this day. In addition to that, the PA dedicated to the robot in the morning is not yet totally busy with this task. Indeed we have only 50 prescriptions to prepare, and at this level the robot does not require a lot of PA time. For example the dispensing robot HEMERA by SINTECO is able to prepare doses for around 100 patients by hour. If one prescription is equal to one patient the PA is only busy for less than 30 minutes and up to 1 hour with this production. Therefore we have a margin of time between 2,5 and 3,5 hours a day and consequently between 12,5 and 17,5 hours a week. Finally, as previously stated a dispensing robot has a big impact on pharmacy production.

CONCLUSION

The contribution of this article is to present some tools to help practitioners in taking strategic and organizational decisions. Although they are dedicated especially to pharmacists they are also usable for other decision-makers.

These tools give them the capacity to build model, to explore new solutions and to know their influence. This is notably the role of our simulation model presented here. Indeed, it evaluates possible configurations of MUP and highlights their consequences. This model is able to do so because it includes a lot of parameters, so it is suitable for different hospitals like the University Hospital of Clermont-Ferrand.

The first tests with the simulation model were successful. The results were close to reality for our hospital and both methodology and results have been

validated by pharmacists. Now it is possible to test and submit some solutions to this hospital. However, there is still some configurations to test and improve (for example unit dose robot) before having a complete simulation model.

Furthermore, a sensibility analysis will be led on parameters (numbers of workers, robot cycles, etc.) to determinate their influence on the system.

Finally, bringing our simulation model in line with the generic processes model will be our next work. In other words, the aim is to automate the translation of BPMN model into SIMIO model. Moreover, the development of a graphical interface for recording and reading results is imperative to make it easier for many practitioners to use the simulation model.

REFERENCES

- Chauvet, J. 2009. *Une méthodologie de modélisation pour les systèmes hospitaliers : application sur le Nouvel Hôpital Estaing*. PhD Thesis. Blaise Pascal University, Clermont-Ferrand, France.
- Gourgand, M. and Kellert, P. 1991. "Conception d'un environnement de modélisation des systèmes de production". In *3ème congrès international de génie industriel* (Tour, France).
- Gourgand, M. and Kellert, P. 1992. "An object-oriented methodology for manufacturing system modeling", *Summer Computer Simulation Conference* (Reno, Nevada, USA). 1123–1128.
- Huet, J.C., Paris, J.L., Gourgand, M. and Kouiss, K. 2010. 'Modèle de connaissance générique du circuit du médicament dans un hôpital'. In *MOSIM'10 : 8^{ème} ENIM IFAC International Conference of Modeling and Simulation* (Hammamet, Tunisia).
- Huet, J.C. 2011. *Proposition d'une méthodologie de réingénierie pour le contrôle par le produit de systèmes manufacturiers : Application au circuit du médicament d'un hôpital*. PhD Thesis. Blaise Pascal University, Clermont-Ferrand, France.
- Rodier, S. 2010. *Une tentative d'unification et de résolution des problèmes de modélisation et d'optimisation dans les systèmes hospitaliers. Application au Nouvel Hôpital Estaing*. PhD Thesis, Blaise Pascal University, Clermont-Ferrand, France.

AUTHOR BIOGRAPHIES

JOHAN ROYER is a PhD student and a hospital engineer in University Hospital of Clermont-Ferrand. He obtained an engineer diploma from IFMA in 2011 specialised in logistics and industrial systems.

JEAN-LUC PARIS is professor at IFMA since 2005. His research areas are modeling, design and optimisation of business process (mainly in healthcare domain) and Knowledge Management.

MICHELLE CHABROL is an associate professor in ISIMA. Her research area is modeling methodology.

A SCOR BASED ANALYSIS OF SIMULATION IN SUPPLY CHAIN MANAGEMENT

Wolfgang Kersten
Muhammad Amad Saeed
Hamburg University of Technology
Institute of Business Logistics and General Management
Hamburg, 21073, Germany
E-mail: logu@tuhh.de

KEYWORDS

Spreadsheet simulation, System dynamics, Discrete-event simulation, Agent-based simulation, Business games, SCOR

ABSTRACT

One of the main goals of simulation in supply chain management is to evaluate the performance and how it can be increased by effectively managing a complex supply chain. Simulation supports managers in decision making at strategic, tactical and operational levels through visualizing, understanding and analyzing the dynamics of the supply chain (SC). This paper provides a detailed analysis of the practice of simulation for different supply chain management (SCM) processes. The supply chain operation reference (SCOR) model is used for the classification of SCM processes. It reports the results of a review and analysis of simulation applications based on literature published within peer-reviewed journals until 2013 in order to provide an up-to-date picture of the role of simulation techniques in SCM. A structured methodology is followed to narrow down the publications (n=569). This research paper mentions different types of simulation in the context to SCM, describes their main characteristics as well as the implementation at different SCM process levels. This leads us to interesting trends and patterns on how different simulation types are applied to different SCM processes in order to answer different managerial questions.

INTRODUCTION

Popular since 1990, SCM has been increasingly widening its scope from the point of origin to the point of consumption (Svensson 2007; Chang and Makatsoris 2000; Lambert and Cooper 2000). According to Makris et al. (2008), “a company’s supply chain comprises geographically dispersed facilities, where raw materials, intermediate products, or finished products are acquired, transformed, stored, or sold and the transportation links that connect facilities along which the products flow”.

The development of SCM has been geared traditionally towards minimization of risks and maximization of profit (Persson and Olhager 2002;

Fawcett et al. 2008; Ashby and Smith 2012). The increase in global competition has also increased the demand for new decision support tools (Almeder et al. 2009). Effective SCM helps to reduce costs and lead times (Tarokh and Golkar 2006) and it requires a carefully defined approach to investigate and evaluate the performance of a SC (Tarokh and Golkar 2006). Thomas and Thomas (2011) distinguished three ways (Mandal 2012) of carrying out SC performance measurement:

- Analytical methods
- Simulation or emulation
- Physical experimentations

In a SCM context, analytical methods such as queuing theory, markov chains, Petri nets, etc., are generally impracticable because of the limited size of the problem they deal with, and many simplifications from the real world case are made in these approaches in order to solve the given problem (Thierry et al. 2008; Almeder et al. 2009; Thomas and Thomas 2011). Physical experiments, such as lab platforms or industrial pilot implementations, suffer technical and cost-related limitations (Mandal 2012). Simulation, relatively often used in comparison to other quantitative models, seems to be the only better approach to model and analyze performance measurement (Kleijnen 2003; Thierry et al. 2008). It allows the design of best decisions in SCM and their evaluation prior to implementation (Maria 1997; Chang and Makatsoris 2000). It makes simulation an excellent tool to reproduce the behavior of complex systems for decision making and can predict the effect of changes to the system, diagnose problems, optimize internal operations and mitigate risks (Tarokh and Golkar 2006; Mandal 2012). Simulation is used to support decision making (Lee et al. 2002; Huan et al. 2004; Tarokh and Golkar 2006) at:

- *The strategic level*, including (re)designing a supply chain, supplier selection, etc.
- *The operational and/or tactical level*, including setting the values of control policies, scheduling, shop floor management, etc.

There are very few articles that provide a comprehensive review about the application of simulation for SCM. One of the earliest reviews by

Kleijnen (2005) explained the use of simulation for SCM and distinguished four simulation types for SCM but did not provide the analysis of the application of different simulation approaches for various SCM processes. Tarokh and Golkar (2006) explained how different simulation types can answer different questions in SCM. Jahangirian et al. (2010) have included a broader range of simulation techniques in manufacturing and business sectors. Tako and Robinson (2012) have given a useful discussion by considering only discrete event simulation and system dynamics in the context of Logistics and SC. Othman and Mustafa (2012) have reviewed different simulation and optimization techniques and analyzed simulation software tools that are being used for SCM, while Mandal (2012) explained the use of simulation for performance measurement in SCM. However, these reviews did not provide a comprehensive conjoint or cross-functional analysis of the application of different simulation approaches for various SCM processes. Therefore, the aim of our literature review is to fill this gap through an extensive coverage of existing academic literature in the field of simulation in SCM and focusing specifically to SCOR-based SCM processes.

REVIEW OF SUPPLY CHAIN SIMULATION METHODS

Computer simulation is the experimentation with a computer model of the system to answer “what-if” questions about the system (Bowden 1998; Seila 2006) and to support decision makers in designing SC operations in order to study their collective and dynamic behaviors (Yancez 2009). In the mid 90’s a large number of software packages, used to support SCM operations, have emerged that are called supply chain management systems (Chang and Makatsoris 2000). From scanning of established literature, five simulations (Kleijnen and Smits 2003; Allwood and Lee 2005; Hao and Shen 2008; Yanez et al. 2009; Jain et al. 2013; Kocaoglu et al. 2013) approaches were distinguished that are used in context of SCM:

- Spreadsheet simulation
- System dynamic
- Discrete-event simulation
- Agent-based simulation
- Business game

A simulation model of a SC can be an entire or standalone model reproducing all nodes, or using more integrated models (one for each node) (Terzia and Cavalierib 2004; Thierry et al. 2008).

Spreadsheet Simulation

“Spreadsheet simulation” refers to the use of a spreadsheet as a platform for representing simulation models and performing simulation experiments (Seila 2006; Othman and Mustafa 2012). It can be used for sampling distribution, test of hypotheses, etc., (Johnson

2011). Spreadsheet simulation is often a simple, economical and relatively straightforward approach. It is possible to develop a simple time slice model but it is difficult to develop a model animation (Kleijnen 2005; Othman and Mustafa 2012; Mishra and Chan 2012). The most prevalent spreadsheet today is Microsoft Excel, which is part of the Microsoft Office package (Greasley 1998; Seila 2006). Any set of calculations in a spreadsheet can be considered as a model. In several studies spreadsheet simulation has not been cited as a formal method of analyzing SCM (Othman and Mustafa 2012) but in a business setting the spreadsheet platform is available to a wide range of decision making (Greasley 1998; Othman and Mustafa 2012) e.g. Koo et al. (1994) used spreadsheet for performance evaluation of a manufacturing system, Sui et al. (2010) used it to determine the replenishment policy in a vendor managed inventory system. Spreadsheet is a tool which can be combined with all other simulation approaches (Thierry et al. 2008).

System Dynamics

According to Labarthe et al. (2007) supply chain modeling and simulation was originally based on system dynamics. System dynamics (SD) is a simulation in which the state of a system varies continuously (Tako and Robinson 2012; Mandal 2012). In system dynamics, firms are viewed as complex systems with different types of flows (e.g. material, orders, manpower, technology, etc.) and stocks (e.g. WIP at a given point in time) (Kleijnen 2005). System dynamics is based on flow models where it is not possible to differentiate between individual elements (Mandal 2012). The managerial control is realized by changing the rate of variables (Kleijnen 2005) (e.g. production rate, sales rate, etc.), which will change the flow as well as the stocks. A SD model takes the feedback principle (closed loop effect) into account which plays a crucial role i.e. managers will take corrective actions if there is an undesirable variation of a targeted value of a performance indicator (Kleijnen 2003; Thierry et al. 2008; Mandal 2012).

Discrete-event Simulation

Discrete event simulation (DES), as the name implies, is a simulation where state changes occur at discrete points in time (Mandal 2012) and is widely used for SC planning (Almeder et al. 2009; Tako and Robinson 2012). A DES is more detailed than the previous two simulation types (Kleijnen & Smits 2003). It represents individual events, e.g. arrival of an individual customer order or the departure of a production lot (Kleijnen 2003). DES is an important method in SCM and used to support decision making for different SCM processes, e.g. Finke et al. (2012) used DES to study changes in production lead time if there are disruptions in operations of an individual processing time of a task. According to Tako and Robinson (2012), hybrid

simulation approaches were used to model SCM processes to support strategic decision making as DES does not represent systems at an aggregate level. Lee et al. (2002) proposed a combined modeling architecture for SC simulation, in which they presented a simple example of a supply chain model dealing at strategic level of a supply chain.

Agent-based Simulation

Over the last few years, agent-based systems are becoming more and more effective tools for solving SCM problems (Mele et al. 2007). So this simulation approach should be added to the above set of simulation approaches for SCM (Eldabi et al. 2008; Yanez et al. 2009). In agent-based simulation, an agent is a real or virtual entity that encapsulates the behaviors of different entities and acts on itself and its surrounding world (Saberli et al. 2012; Ilie-Zudor and Monostori 2009). The multi-agent system (MAS) collaborates, and considers exchange of information and relationships among other agents in order to obtain improved solutions (Nikolopoulou and Ierapetritou 2012). Supply chain activities such as sourcing, planning, delivery and their interactions are represented by different agents in multi-agent systems (Saberli et al. 2012). Agents are autonomous, reactive, and proactive and have social ability (Julka et al. 2002). Due to these advantages, MAS have become a promising tool for solving SCM problems in the last few years (Puigjaner and Gosalbez 2008).

Business Games

Modeling of human behavior is more difficult in comparison to modeling of different SCM processes (Kleijnen 2005). It can be achieved by letting managers operate within a simulated world (consisting of SC and its environment) (Kleijnen and Smits 2003; Mandal 2012). Business games are used for educational and research goals (Kleijnen 2003) and have no relationships to game theory which is applied to SC (Mandal 2012). The beer game is widely used among several proposed games (Mandal 2012). Kleijnen and Smits (2003) distinguished business games in two subtypes:

- i. In *Strategic games* several teams of players (representing companies) compete with each other. Players interact with the simulation model for a fixed number of rounds, like beer game which is used to illustrate the bullwhip effect (Kleijnen 2005; Thierry et al. 2008).
- ii. In the simulated world of *Operational games*, a single team (one or more players at a time) interacts with a simulation model for several rounds, like games for training of production scheduling (Kleijnen 2003; Thierry et al. 2008).

Furthermore, based on complexity level, Thierry et al. (2008) categorized business games as board games

(simple) that are played with tokens on specific boards and sophisticated games (more realistic) that run on computer devices.

The type of simulation applied in SCM depends on the problem that needs to be solved e.g. use of SD demonstrates the bullwhip effect, DES simulation can quantify fill rates, use of business games to educate and train users (Kleijnen 2003).

CLASSIFICATION OF SUPPLY CHAIN MANAGEMENT PROCESSES BASED ON SCOR MODEL

Every business process has different characteristics. Thus a good understanding of the SCM processes is necessary before developing a simulation model (Chang and Makatsoris 2000). In the past, different metrics were used to measure performance at different SC levels and models for decisions making at strategic level are scarce (Huan et al. 2004). It is not achievable to have a model that perfectly represent SCM but a closely adapted model (Irfan et al. 2008; Mandal 2012) i.e. SCOR. In 1996, the Supply Chain Council (SCC) has developed and released a structured framework model 'SCOR' for SCM systems and practices (Li et al. 2011). SCOR model is the first cross-industry framework for evaluating and improving SC performance and management (Stewart 1997; Huan et al. 2004). The benefit of the 'SCOR' model is that it provides a standard format and a comprehensive methodology to facilitate communication and to improve SC operations (Irfan et al. 2008). It is a flexible framework and common language that can help upper management of a firm to design and reconfigure its SC to achieve the desired performance both internally and externally (Huan et al. 2004). SCOR is widely used in academia and practice. For these reasons we use the SCOR framework as a base of our analysis.

'SCOR' is a business process framework that spans from the supplier's supplier to the customer's customer ('SCOR' 2012). 'SCOR 11.0' describes a SC by four levels of details. *Level 1* is a top level that defines the scope and high level configuration by six core processes, i.e. plan, source, make, deliver, return and enable. *Level 2* is a configuration level and processes at level 2, along with their positioning, determine the SC strategy. *Level 3* is a process element level and describes the steps that need to be performed to execute all of the level 2 processes. *Level 4* is the implementation level and describes industry specific activities which are required to perform level 3 processes.

Each of the SCOR components is considered as an important intra-organizational function and a critical inter-organization process (Erkan and Bac 2011). SCOR core processes are defined below and shown in figure 1 (SCOR 2012). The process reference model utilized in our research is based on 'SCOR 11.0'. Furthermore, we

concentrated the research scope to Level 1 process types specifically.

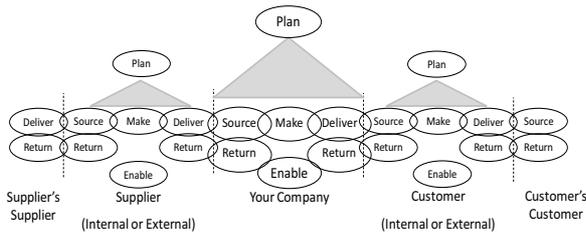


Figure 1: SCOR Model

RESEARCH METHODOLOGY

The aim of this study is to analyze the use of simulation in SCM; looking specifically into the detail of SCM processes are simulated. In order to accomplish the goal our analysis is based on the frequency with which SCM processes are simulated using spreadsheet simulation, system dynamics, discrete event simulation, agent-based simulation and business games. The study is based on a review of journal articles which describe the application of simulation techniques to different processes in SCM. Following two research questions are addressed here:

- Which simulation approach is used to a great extent across different SCOR-based SCM processes?
- Which SCOR-based SCM process widely employs simulation approach to support decision making?

Our expectations established on initial literature analysis, that all of above considered simulation techniques are used to model various processes of SCM. The systematic literature review undertaken follows two stages:

- i. Identification of journal articles and applied simulation approaches
- ii. Classification of journal articles by SCOR-based SCM processes

A detailed diagram of steps taken while selecting articles is described in figure 2 (adopted from Eldabi et al. 2008).

Identification of Journal Articles and Adopted Simulation Approaches

A list of keywords was generated based on an initial literature analysis. Later, these keywords were used to find related scientific articles using Science direct, EBSCO host and Web of Science databases (DB). This provided a multidisciplinary collection of literature, but journal articles that report simulation and are relevant to SCM were selected. Science direct is a leading scientific database offering more than 2,200 journals and almost 26,000 books titles (Elsevier 2014). EBSCO host offers more than 375 full-text and secondary research databases plus subscription management services for

355,000 e-Journals and e-Journal packages (Ebscohost 2014). Web of Science provides access to the world's leading citation databases and covers over 12,000 of the highest impact journals worldwide (WEB of Science 2014). All three databases contain articles from top ranked journals in the field of SCM.

Our initial set of keywords consisted of spreadsheet simulation, system dynamics, discrete event simulation, business games and agent-based simulation. Each of these keywords were combined with 'supply chain' as a second keyword by using 'and' as an operator, and searched inside selected scientific databases. Based on the initial results and a review, more keywords were generated and searched with full-text option in each database. Results from each database were compiled in three different Microsoft Excel spreadsheets. It comprised articles published until September, 2013 and included 4 articles that will be published in 2014 but are already available online.

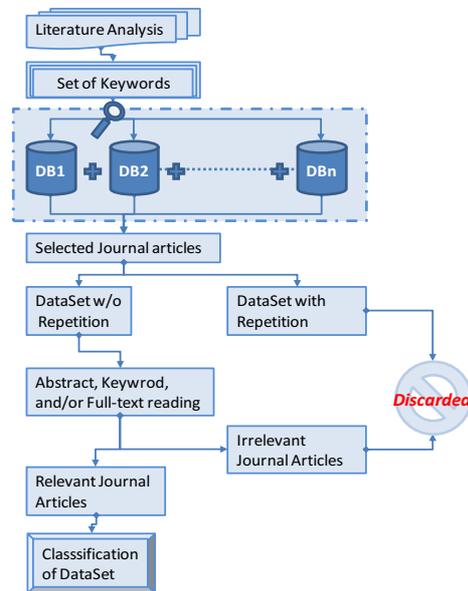


Figure 2: A framework for Literature review

After combining all of the results from three spreadsheet datasets, duplications were noticed. That happened because some of the publications were cited in more than just one database as well as because of the application of different combination of keywords. After removing all duplicates, the initial dataset resulted in 4200 publications. The search was limited to peer-reviewed publications from scientific journals only and no books, conference papers, magazines, etc., were considered. The applied procedure was to read the abstract and/or also full article if the topic was not clear from the title and author keywords. After screening, a reduced list of 569 publications remained. Each of the selected articles has adopted one of the above mentioned simulation (i.e. spreadsheet simulation, system dynamics, discrete-event simulation, agent-based

simulation, business games) approaches or hybrid (i.e. combination of two or more) simulation.

Classification of Journal articles by SCOR-based SCM Processes

After keeping only unique and relevant publications in the Excel spreadsheet, the next step is to classify each article into SCM processes. Different authors classify SCM differently as it is a vast field, covering a variety of topics (Tako and Robinson 2012). For this purpose, we adopted a classification of SCM processes based on the ‘SCOR’ model. All 569 selected journal articles from step one have been classified for further analysis. Sometimes authors discussed more than one process or issue in their research work; in that case publications were classified accordingly.

RESULTS & ANALYSIS

Journal articles classified in our Excel dataset were analyzed to address the research questions mentioned in the research methodology and the results were presented with various perspectives, as given below in detail.

Distribution by Publication Year, Document type and Language

At first, only English language journal articles were selected for conducting this study. Figure 3 shows the distribution of journal publications by year, and Table 1 shows the percentage and the number of publications per year. One of the interesting findings is the increase in SC simulation between the years 2000 and 2010. Since it is stagnating on a high level and before 2000 there was only small number of publications. Based on the number of publications per year, it can be stated that simulation techniques are becoming ever more important. The percentage of number of publications for one year is calculated as a portion of the publications identified for the purpose of analysis.

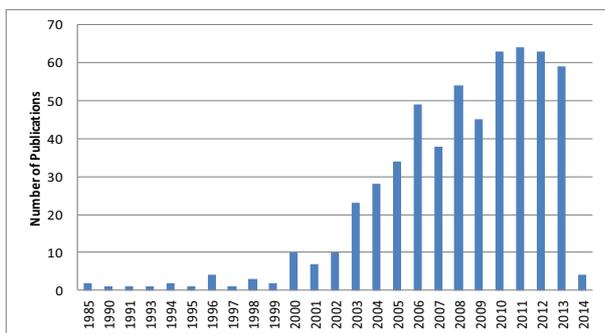


Figure 3: Distribution of Publications per Year

Distribution of publications by Source Title

Our dataset contains publications from different journals. Table 2 shows a list of most cited journals in our selected dataset. “International Journal of Production Research”, “International Journal of

Production Economics” and “European Journal of Operational Research” are reported as top three journals with the most publications.

Table 1 : Percentage of Publications per Year

Pub. Year	#	%	Pub. Year	#	%
1985	2	0,35	2003	23	4,04
1990	1	0,18	2004	28	4,92
1991	1	0,18	2005	34	5,98
1993	1	0,18	2006	49	8,61
1994	2	0,35	2007	38	6,68
1995	1	0,18	2008	54	9,49
1996	4	0,70	2009	45	7,91
1997	1	0,18	2010	63	11,07
1998	3	0,53	2011	64	11,25
1999	2	0,35	2012	63	11,07
2000	10	1,76	2013	59	10,37
2001	7	1,23	2014	4	0,70
2002	10	1,76	Total	569	100,0

Table 2: List of Top Cited Journals in dataset

International Journal	# of Publications
International Journal Of Production Research	83
International Journal Of Production Economics	63
European Journal Of Operational Research	25
International Journal Of Computer Integrated Manufacturing	21
Computers & Industrial Engineering	16
Decision Support Systems	15
Expert Systems With Applications	15
Simulation Modelling Practice And Theory	13
Computers & Chemical Engineering	12
Computers In Industry	12

Distribution by SCOR-based SCM Processes

One of the aims was to determine which objectives of SCM are achieved by simulation and its applicability in supporting decision making at different SCM processes. From figure 4 it can be seen that out of 569 articles, 298 (52%) publications applied simulation for the ‘plan’ process and similarly for the ‘source’ process 24 (4%), for the ‘make’ process 101 (18%), for the ‘deliver’ process 32 (6%), for the ‘return’ process 8 (1%), for the ‘enable’ process 59 (10%) and 47 (8%) publications dealt with more than process at a time.

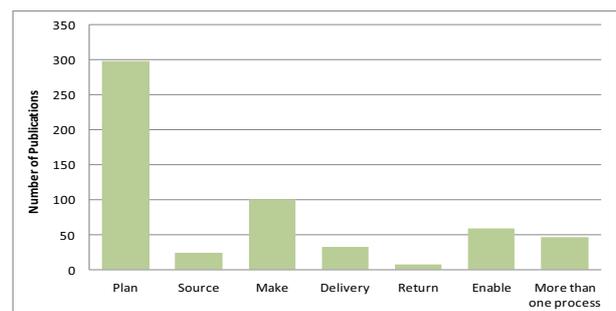


Figure 4: SCOR-based SCM Processes

Distribution by Application of Simulation Approach

Another objective of the study was to determine which simulation approach is widely used by researchers to support decision making in SCM. According to figure 5, out of 569 articles, 27% of the publications applied DES approach, 25% of the publications applied agent-based simulation, 19% of the publications applied SD, 15% of the publications applied the spreadsheet simulation, 9% of the publications applied hybrid simulation and the remaining 5% used business games.

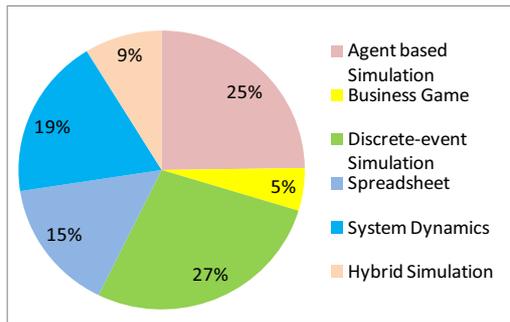


Figure 5: Simulation in SCM

Distribution based on Simulation Approach and SCM Process

The use of simulation approaches across the different SCM processes can be analyzed from two different perspectives:

- From the SCM process
- From simulation approach

Accordingly the following questions arise:

1. What are the main simulation approaches that support a certain SCOR-based SCM process? (see table 3)

It can be seen that for solving the 'plan' process problems, agent-based simulation is used i.e. 26% out of

298 publications, following are DES and SD i.e. 21% and 20% respectively. For dealing issues within 'source' processes, spreadsheet simulation is widely used i.e. 46% out of 24 publications. For 'make' processes, DES is mostly used i.e. 45% following agent-based simulation i.e. 24%. For 'deliver' processes, again DES is used more frequently i.e. 50% following spreadsheet simulation i.e. 22% of the 32 publications. For SCM 'enable' processes, agent-based simulation and SD are used for 32% and 29% respectively. Similarly, for 'return' processes SD is used most frequently i.e. 65% out of 8 publications.

2. What are the major fields of application (i.e. SCOR SCM processes) for a specific simulation approach? (see table 4)

Agent-based simulation is used 55% (77) for 'plan' processes out of 141 publications, following 'make' processes i.e. 17% (24). DES is also used most frequently to simulate 'plan' processes i.e. 39% out of 157 publications and 29% for 'make' processes. Similarly, SD is used for the 'plan' processes for 58% out of 106 publications. Spreadsheet simulation is used for 46% in 'plan' processes out of 87 publications and business games are used to simulate 'plan' processes for 79% publications out of 28. Hybrid simulation is largely applied for 'plan' processes, i.e. 72% out of 50 publications.

Implementation of Simulation Approaches with Time

According to Gartner (2013), there is an increase of 8% in sales of SCM software from 2008 to 2012. From our Excel dataset, as represented in figure 6, it can be interpreted that a sharp increase in use of agent-based simulation is noted in recent years, and a constant trend in use of business games and spreadsheet simulation techniques is noticed. Apart from frequency, simulation in SCM is gaining much more popularity than in the past few years and a continuous increase in all types of simulation approaches is noticed during the analysis.

Table 3: Main Simulation Approaches to Support a certain SCOR-based SCM process

Simulation Approach	Plan		Source		Make		Deliver		Return		Enable		More than one process	
	Publications	%age	Publications	%age										
Spreadsheet	40	13%	11	46%	12	12%	7	22%	1	13%	6	10%	10	21%
System Dynamics	61	20%	1	4%	14	14%	2	6%	5	63%	10	17%	13	28%
Discrete-event Simulation	62	21%	4	17%	45	45%	16	50%	0	0%	17	29%	13	28%
Agent based Simulation	77	26%	8	33%	24	24%	5	16%	0	0%	19	32%	8	17%
Business Game	22	7%	0	0%	0	0%	0	0%	1	13%	4	7%	1	2%
Hybrid Simulation	36	12%	0	0%	6	6%	2	6%	1	13%	3	5%	2	4%
Total	298	100%	24	100%	101	100%	32	100%	8	100%	59	100%	47	100%

Table 4: Major Fields of Application for a Specific Simulation Approach

SCM Processes	Spreadsheet		System Dynamics		Discrete-event Simulation		Agent based Simulation		Business Game		Hybrid Simulation	
	Publications	%age	Publications	%age	Publications	%age	Publications	%age	Publications	%age	Publications	%age
Plan	40	46%	61	58%	62	39%	77	55%	22	79%	36	72%
Source	11	13%	1	1%	4	3%	8	6%	0	0%	0	0%
Make	12	14%	14	13%	45	29%	24	17%	0	0%	6	12%
Deliver	7	8%	2	2%	16	10%	5	4%	0	0%	2	4%
Return	1	1%	5	5%	0	0%	0	0%	1	4%	1	2%
Enable	6	7%	10	9%	17	11%	19	13%	4	14%	3	6%
More than One SCM process	10	11%	13	12%	13	8%	8	6%	1	4%	2	4%
Total	87	100%	106	100%	157	100%	141	100%	28	100%	50	100%

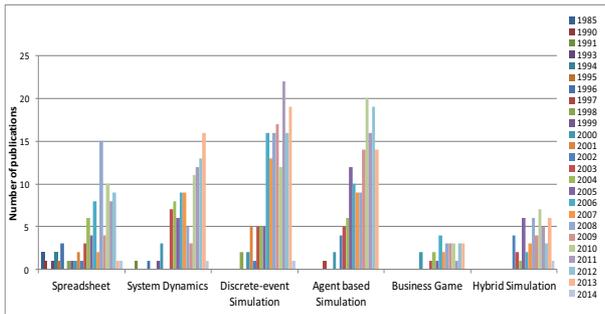


Figure 6: Implementation of Simulation Approaches with Time

CONCLUSION AND FURTHER RESEARCH

Data from three scientific databases were collected. It was experienced through database search that an inclusion of new scientific databases will not distort the result and will lead only to a few new entries in our dataset. An increasing number of repetitions were noticed while searching the second database after EBSCO host and after that more repetitions were noticed while searching for the third database i.e. Web of Science (see figure 7). So, it was decided to limit database search to a maximum of three scientific databases.

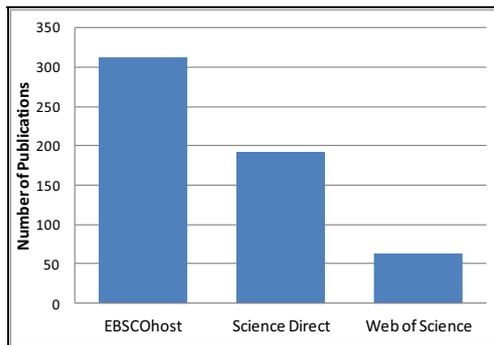


Figure 7: Distribution of Publications in Dataset

All simulation approaches are used for solving problems and support the decision making at almost all SCOR-based SCM processes. There are however, different degrees of use across SCM, which suggests some preference for one approach over the others.

The analysis of journal articles reflects that DES is more frequently used in comparison to other simulation types and at the same time application of agent-based simulation in SCM is growing at a faster rate than other simulation approaches. English language journal articles until 2013 were reviewed in order to elucidate the two questions initially placed.

The first question, frequency of use of simulation approaches for SCM processes? Use of a particular type of simulation depends on the type of the SCM question to be answered. DES approach has been applied most frequently in comparison to other simulation approaches. It can be seen from the analysis that all

simulation approaches are getting popularity in academia.

The second question, which SCOR-based SCM process largely employs simulation as a decision support tool? The research shows that the ‘plan’ process employs simulation most frequently in comparison to other processes in SCM.

From the percentage use of simulation for SCM processes, it is revealed that for ‘plan’ process of SCOR model, agent-based simulation was mostly used and the use of all other simulation approaches is relatively small. For the ‘source’ process the spreadsheet simulation is widely used. For ‘make’ and ‘deliver’ processes, DES approach is mostly used and the ‘return’ process employs the SD approach as a decision support tool in SCM. There are only 9% of the publications that uses more than just one simulation approach at a time. So, it can be expected that a hybrid simulation approach could be an area of a future research and development.

This study makes a contribution to limited knowledge about the use of different simulation approaches in the context of SCM. Future work could expand the focus specifically on phenomena in SCM such as SC risk management, SC complexity, SC sustainability, etc.

Further research could also focus and analyze more details of the ‘SCOR’ model e.g. going deeper to SCOR’s Level 2 and Level 3 processes. This will help to find out how far the different types of simulation are applied in certain SCM processes and contexts.

The findings of this study are based only on journal publications. These publications show more the academic interest than the practice use of simulation (Tako and Robinson 2012). This may affect the analysis results which are presented in our study as it might not reflect the full range and frequency of use of simulation approaches in the SCM practice: Thus future work could extend on this study by considering more practice-oriented journals and also other literature resources e.g. conference papers and/or book chapters, etc, to conduct a similar review about the application of simulation in SCM with a more practiced based focus.

This paper does not provide information about the selection and the success of a specific simulation approach for a particular SC process. Such study can also lead to further comparison studies.

REFERENCES

- Allwood, J.M. and J.H. Lee. 2005. “The Design of an Agent for Modeling Supply Chain Network Dynamic”. *International Journal of Production Research* 43, No. 22(Nov), 4875–4898.
- Almader, C.; M. Preusser; and R.F. Hartl. 2009. “Simulation and Optimization of Supply Chains: Alternative or Complementary Approaches?” *OR Spectrum* 31, 95–119.
- Ashby, A.; M. Leat and M.H. Smith. 2012.” Making Connections: A Review of Supply Chain Management and

- Sustainability Literature". *Supply Chain Management: an International Journal* 17(5) 497–516.
- Bowden, R. 1998. *The Spectrum of Simulation Software III Solutions* 44-54 (May)
- Chang, Y. and H. Makatsoris. 2000. "Supply Chain Modeling Using Simulation" *I. J. Of SIMULATION* 2, No1, 24-30.
- Datta, P. and C. Martin. 2009. "Information Sharing and Coordination Mechanisms for Managing Uncertainty in Supply Chains: A Simulation Study" *International Journal of Production Research* 49(3), 765-803.
- Ebscohost. 2014. "About EBSCO".
<http://www.ebsco.com/about>
- Eldabi, T.; M. Jahangirian; A. Naseer; L.K. Stergioulas; T. Young and N. Mustafee. 2008. "A Survey of Simulation Techniques in Commerce and Defence". *Proceedings of the OR Society 4th Simulation Workshop (SWO8)*, 275-284.
- Elsevier. 2014. "Who Uses ScienceDirect".
<http://www.elsevier.com/online-tools/sciencedirect/who-uses-sciencedirect>
- Erkan, T.E. and U. Bac. 2011. "Supply Chain Performance Measurement: A Case Study about Applicability of SCOR Model in a Manufacturing Industry Firm" *International Journal of Business and Management Studies* 3, No.1, 381-390.
- Fawcett, S.E. and G.M. Magnan. 2008. "Benefits, Barriers, and Bridges to Effective Supply Chain Management" *Supply Chain Management: an International Journal* 13, No.1, 35–48.
- Finke, G.R.; M. Singh and P. Schönsleben. 2012. "Production Lead Time Variability Simulation-Insights from A Case Study" *International Journal of Industrial Engineering* 19, No.5, 213.
- Gartner. 2013. *Global Supply Chain Management (SCM) Software Market Revenue from 2008 to 2012*.
- Greasley, A. 1998. "An Example of a Discrete-Event Simulation on a Spreadsheet" *SIMULATION* 70, 148.
- Hao, Q. and W. Shen. 2008. "Implementing a Hybrid Simulation Model for a Kanban-Based Material Handling System". *Robotics and Computer-Integrated Manufacturing* 24, 635–646.
- Huan, S.H.; S.K. Sheoran and G. Wang. 2004. "A Review and Analysis of Supply Chain Operations Reference (SCOR) Model". *Supply Chain Management: An International Journal* 9(1), 23-29.
- Ilie-Zudor, E. and L. Monostori. 2009. "Agent-Based Framework for Pre-Contractual Evaluation of Participants in Project-Delivery Supply-Chains". *Assembly Automation* 29(2), 137–153.
- Irfan, D.; X. Xiaofei and D.S. Chun. 2008. "A SCOR Reference Model of the Supply Chain Management System in an Enterprise". *The International Arab Journal of Information Technology* 5(3), 288-295
- Jahangirian, M.; T. Eldabi; A. Naseer; L.K. Stergioulas and T. Young. 2010. "Simulation in manufacturing and business: A review". *European Journal of Operational Research* 203, 1-13.
- Jain, S.; E. Lindskog, J. Andersson and B. Johansson. 2013. "A Hierarchical Approach for Evaluating Energy Trade-offs in Supply Chains". *International Journal of Production Economics* 146, 411–422.
- Johnson, A.C. 2011. "Introducing Simulation via the Theory of Records" *Decision Sciences Journal of Innovative Education* 9(3), 411-419.
- Julka, N.; R. Srinivasan and I. Karimi. 2002. "Agent-Based Supply Chain Management-I: Framework" *Computers and Chemical Engineering* 26, 1755-1769.
- Kleijnen, J.P.C. and M.T. Smits. 2003. "Performance Metrics in Supply Chain Management". *Journal of the Operational Research Society* 54,507-514.
- Kleijnen, J.P.C. 2003. "Supply Chain Simulation: A Survey" Discussion Paper 2003-103 Tilburg University.
- Kleijnen, J.P.C. 2005. "Supply Chain Simulation Tools and Techniques: A Survey" *International Journal of Simulation & Process Modeling*, 1 (1/2), 82-89.
- Kocaoglu, B.; B. Gülsün and M. Tanya. 2013. "A SCOR-based Approach for Measuring a Benchmarkable Supply Chain Performance" *Journal of Intelligent Manufacturing* 24, 113–132.
- Koo, P.H.; C.L. Moodie and J.J. Talavage .1994. "Performance Evaluation of Manufacturing Systems: A Spreadsheet Model". *Computers & Industrial Engineering* 673-688.
- Labarthe, O.; B. Espinasse; A. Ferrarini and B. Montreuil. 2007. "Toward a Methodological Framework for Agent-Based Modeling and Simulation of Supply Chains in a Mass Customization Context". *Simulation Modeling Practice and Theory* 15,113–136.
- Lambert, D.M. and M.C. Cooper. 2000. "Issues in Supply Chain Management". *Industrial Marketing Management* 29, 65–83.
- Lee, Y.H.; M.K. Cho and Y.B. Kim. 2002. "Supply Chain Simulation with Discrete-Continuous Combined Modeling". *Computers & Industrial Engineering* 43, 375-392.
- Lia, L.; Q. Subc and X. Chen. 2011. "Ensuring Supply Chain Quality Performance through Applying the SCOR Model". *International Journal of Production Research* 49(1), 33–57.
- Makris, S.; P. Zoupas and G. Chryssolouris. 2011. "Supply Chain Control Logic for Enabling Adaptability under Uncertainty". *International Journal of Production Research* 49(1), 121–137.
- Mandal, S. 2012. "Reviewing Simulation in Supply Chain Management". *International Journal of Research in Management, Economics and Commerce* 2 No.11 (Nov), 412-420.
- Maria, A. 1997. "Introduction to Modeling and Simulation". *Winter Simulation Conference* 7-13.
- Makris, S.; V. Xanthakis; D. Mourtzis and G. Chryssolouris. 2008. "On the Information Modeling for the Electronic Operation of Supply Chains: A Maritime Case Study". *Robotics and Computer-Integrated Manufacturing* 24, 140–149.
- Márquez, A.C. 2010. *Dynamic Modeling for Supply Chain Management Dealing with Front-end, Back-end and Integration Issues*. Springer-Verlag London Limited.
- Mele, F.D.; G. Guillen; A. Espuna and L. Puigjaner. 2007. "An Agent-based Approach for Supply Chain Retrofitting under Uncertainty" *Computers and Chemical Engineering* 31, 722–735.
- Mishra, M. and F.T.S. Chan. 2012. "Impact Evaluation of Supply Chain Initiatives: A System Simulation Methodology". *International Journal of Production Research* 50 No. 6, 1554–1567.
- Nikolopoulou, A. and M.G. Ierapetritou. 2012. "Hybrid Simulation based Optimization Approach for Supply Chain Management". *Computers and Chemical Engineering* 47, 183– 193.

- Othman, S.N. and N.H. Mustaffa. 2012. "Supply Chain Simulation and Optimization Methods: An Overview". *Third International Conference on Intelligent Systems Modeling and Simulation* 161-167.
- Persson, F. and J. Olhager. 2002. "Performance Simulation of Supply Chain Designs". *International Journal of Production Economics* 77, 231-245.
- Puigjaner, L. and G.G. Gosalbez. 2008. "Towards an Integrated Framework for Supply Chain Management in the Batch Chemical Process Industry". *Computers and Chemical Engineering* 32 No. 4-5(Apr), 650-670.
- Saberi, S.; A.S. Nookabadi and S.R. Hejazi. 2012. "Applying Agent-based System and Negotiation Mechanism in Improvement of Inventory Management and Customer Order Fulfillment in Multi Echelon Supply Chain". *Arabian Journal for Science & Engineering* 37,851-861.
- SCOR. 2012. "Supply Chain Operations Reference Model". *Version 11.0*.
- SCOR 11.0 <https://supply-chain.org/f/SCOR11QRG.pdf>
- Seila, A.F. 2006. "Spreadsheet Simulation". *Proceedings of the 2006 Winter Simulation Conference* 11-18.
- Stewart, G. 1997. "Supply-Chain Operations Reference Model (SCOR): The First Cross-Industry Framework for Integrated Supply-Chain Management". *Logistics Information Management* 10 No. 2, 62-67.
- Sui, Z.; A. Gosavi and L. Lin. 2010. "A Reinforcement Learning Approach for Inventory Replenishment in Vendor-Managed Inventory Systems with Consignment Inventory". *Engineering Management Journal* 22 No. 4, 44-53.
- Svensson, G. 2007. "Aspects of Sustainable Supply Chain Management (SSCM): Conceptual Framework and Empirical Example". *Supply Chain Management: An International Journal* 12 No.4 262-266.
- Tako, A.A. and S. Robinson. 2012. "The Application of Discrete Event Simulation and System Dynamics in the Logistics and Supply Chain Context". *Decision Support Systems* 52, 802-815.
- Tarokh, M. J. and M. Golkar. 2006. "Supply Chain Simulation Methods". *Service Operations and Logistics, and Informatics* 448-454.
- Terzia, S. and S. Cavalierib. 2004. "Simulation in the Supply Chain Context: A Survey". *Computers in Industry* 53 No. 1, 3-16.
- Thierry, C.; A. Thomas and G. Bel. 2008. "Supply Chain Management Simulation: An Overview". In *Simulation for Supply Chain Management*, Thierry, C.; A. Thomas and G. Bel (Eds.). Wiley-ISTE, 1-36.
- Thomas, P. and A. Thomas. 2011. "Multilayer Perception for Simulation Models Reduction: Application to a Sawmill Workshop". *Engineering Applications of Artificial Intelligence* 24 No.4, 646-657.
- Volling, T. and T.S. Spengler. 2011. "Modeling and Simulation of Order-driven Planning Policies in Build-to-order Automobile Production". *International Journal of Production Economics* 131, 183-193.
- WEB of Science. 2014. "Fact: A True Citation Index". In *Real Facts of the Citation Connection*. <http://wokinfo.com/citationconnection/realfacts/#truecitindex>
- Yanez, F.C.; J.M. Frayret; F. Leger and A. Rousseau. 2009. "Agent-Based Simulation and Analysis of Demand-Driven Production Strategies in the Lumber Industry". *International Journal of Production Research* 47 No. 22, 6295-6319.

AUTHOR BIOGRAPHIES

Prof. Dr. Dr. h.c. Wolfgang Kersten (doctoral degree from the University of Passau) is Professor for Logistics and Director of the Institute of Business Logistics and General Management at the Hamburg University of Technology. Prof. Kersten's research focuses on the areas of supply chain risk and complexity management, sustainability in supply chain management, project management, and supply chain design and simulation. His work appears in several anthologies, journals and conference proceedings. His e-mail address is: logu@tuhh.de

Muhammad Amad Saeed (M.Sc. degree from Hamburg University of Technology) is research associate at the Institute of Business Logistics and General Management at Hamburg University of Technology. His research focuses on simulation in supply chain management and supply chain sustainability. His e-mail address is: muhammad.saeed@tuhh.de

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems

-

Special Session

Implementation of the genetic algorithm by means of CUDA technology involved in travelling salesman problem

Anna Plichta

Tomasz Gąciarz

Cracow University of Technology, Department of Computer Science

31-155 Cracow, ul. Warszawska 24

Email: aplichta@pk.edu.pl

Email: tga@pk.edu.pl

Bartosz Baranowski

Email: bbaranowski@onet.pl

Szymon Szomiński

AGH University of Science and Technology

30-059 Cracow, Al. A. Mickiewicza 30

Email: szsz@agh.edu.pl

KEYWORDS

genetic algorithm, CUDA technology, travelling salesman problem

ABSTRACT

The research was intended to solve the travelling salesman problem by means of genetic algorithms. The implementation of the algorithm was by virtue of CUDA technology. The research was focused on checking how much the system can improve if instead of classical CPU processors one uses GPU graphical processors enabled to perform the operations parallel. The algorithm was implemented in the high level CUDA C language. Thus, measuring the pure time of performance of the algorithm could be the single but reliable point of comparison between two above mentioned types of processors. Making some operations mutually independent and using CUDA technology makes the task much faster to execute. Due to it complex issues can be solved in a shorter time.

INTRODUCTION

Video cards were primarily used to display graphics, but it was quickly observed that GPU processors may be much faster than CPU on one condition: the problems to sort out should be parallel. Hence, more and more supercomputers of the TOP500 list are based on NVIDIA Tesla video cards and many so called GPGPU technologies were developed (General-Purpose Computing on Graphics Processing Units) which enable for making general computations by virtue of video card processor. The most popular of them are NVIDIA CUDA (Compute Unified Device Architecture), OpenCL (Open Computing Language) and Microsoft DirectCompute.

The aim of the project was to verify CUDA technology and GPU graphic processors in respect of parallel genetic algorithms. Therefore, the genetic algorithm was implemented into the traditional CPU processor-based technology and into GPU graphic processor-based technology in order to compare the efficiency of both implementations. The travelling salesman problem was chosen to check the genetic algorithm.

GENETIC ALGORITHM

Genetic algorithm (GA) by John Holland and David E. Goldberg [4], [2] is the adaptation searching procedure based on the mechanisms of natural selection and natural genetics. These parallel algorithms are simple, general and efficient. Genetic algorithms (GA) can be applied to designing electric machines and circuits, to optimizing routes, to lofting in telecommunication or in computer games. Genetic algorithms can be also used to solve NP problems such as travelling salesman problem [7], [8], [9]. Travelling salesman problem (TSP) pertains to the graph theory and it consists in finding minimal Hamilton cycle in the full weighted graph. Usually, the issue is represented from the point of view of the travelling salesman who has to visit N cities (each city once). The distances between all cities are known. The problem is NP-difficult [16].

As for GA, points belonging to the search space are represented as limited-length chains of limited-alphabet symbols. In order to solve the problem GA (in each generation) transforms the set of points of search space (called chromosomes). That particular collection is called population.

Chromosomes within the particular population compete one another for survival and possibility of reproduction. During each iterative step of the genetic algorithm called genera-

tion all individuals are evaluated and undergo recombination according to their level of usefulness and some "genetic operators". As a result, the following populations are growing better and better. That process is usually suppressed when the desired number of populations is achieved or when the population ceases to change [10].

The function evaluating each chromosome which is called by biologists the fitness function is a measure of benefit or profit we want to maximize. The first genetic operator called reproduction works according to the values of the above-mentioned function. During that process some chromosomes are selected to undergo the next operation (the more adapted the chromosome the higher its chance for selection). All selected chromosomes are coupled randomly. During the crossover process (the second genetic operator) they exchange some elements in order to create new offspring. The third genetic operator called mutation is used in order to avoid premature convergence to one of the local minima and increase the diversity of the population. It changes the values of some content of the chromosome at random consequently making up its almost identical copy [2].

In order to solve the problem with the genetic algorithm, the latter should be represented as chromosome. Moreover, the crossover and mutation should be prepared as well as the fitness function which measures the usefulness of the chromosome regarding the issue.

GA AND THE TRAVELLING SALESMAN PROBLEM

Encoding of the chromosome

For the purpose of travelling salesman problem where each individual represents one route the most common encoding is encoding as integers. Each city has its unique number and appears in the chromosome only once. In the Figure 1 the route starts in the point (city) 3 and leads through points 1,2,5 and 4. In many cases the route is additionally represented as a loop. The additional edge from the point 4 to 3 appears [5].

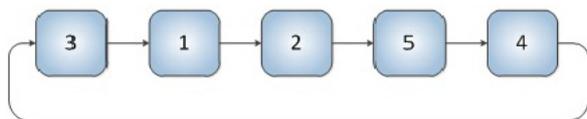


Fig. 1. The example of encoding the chromosome with integers

Genetic operators - selection, crossover, mutation

Two methods of selection were implemented and tested, namely, the method called roulette wheel and the tournament one. Practically, the latter proved to be more efficient.

One should modify the standard crossover procedure so that to take account of the structure of the chromosome adapted specifically to the travelling salesman problem. Three crossover methods were implemented and tested: one-point crossover, two-point crossover and greedy crossover.

However, because in the travelling salesman problem each gene is represented by the unique integer which cannot appear

in the individual more than once, one should add some correcting mechanisms. Just as in the binary representation the first step is to copy the left side of the first parent into the offspring. Then only these genes from the right side of the second parent are copied which the offspring lacks. Finally, one should find and copy the lacking genes into the offspring. For instance, the second parent can be searched from its left side until a lacking gene is found.

Two-point crossover is more advanced version of the one-point crossover. In two-point crossover both parents are cut in two random points in order to get left, right and the middle part. The offspring is joined together from the middle section of one parent and the extreme parts of the second parent. Just as in one-point crossover one can simultaneously create two descendants and the correcting mechanisms should be added to prevent the descendant from having two or more identical genes [19].

First of all, the middle part of the first parent is copied into the descendant. Then, the genes which the descendant lacks are copied from the extreme parts of the second parent. At last, the descendant is given the genes which it still lacks. For instance, the second parent can be searched from its left side until a lacking gene is found.

Greedy crossover is the method applied specifically to the travelling salesman problem (or similar ones) as it can be used only in case each gene in the chromosome in unique and appears only once within it [1].

For the purpose of the travelling salesman problem mutation is usually changing the places of two random gene of the descendant. The occurrence of the mutation is infrequent (its probability is about 0,5%).

The aim function

In the project we applied simple adaptation function consisting in computing the length of the route of the individual. The better (shorter) the route the lesser the value of the adaptation function. Distances between the cities are stored in two-dimension table.

Aiding mechanisms

2opt Heuristics

In order to improve the results in many versions of genetic algorithms one uses also heuristic methods [21]. For instance, 2opt method is often used to solve the travelling salesman problem. It is a method of local optimization consisting in reconfiguring two edges so that to achieve locally shorter route.

IMPLEMENTATION INTO CPU PROCESSOR

To implement the algorithm into the CPU processor the Code::Blocks environment was used together with MinGW compiler (v.4.6.1) based on GCC. In order to provide multithreading the OpenMP library was used.

The C language was chosen mainly because such choice facilitates the implementation of the algorithm into graphic card in CUDA C language in the second stage of the research. The implementation of the algorithm underwent many changes. Initially, it was a simple genetic algorithm provided with one-point crossover and mutation. The selection method was roulette wheel.

In the next version the selection method was replaced with the tournament one and one-point crossover was replaced with two-point crossover. In the next version greedy crossover appeared and additional heuristic methods which improved achieving good results faster. At last, the island model was provided.

The final version of the algorithm was comprised of tournament selection, greedy crossover and standard mutation. The results were also improved by means of the 2opt method. Multithreading was achieved thanks to adding islands served by separate threads. In each island there is a separate instance of the genetic algorithm, served by the separate plot. Instances communicates one another by means of fixed-frequency migration. Each island sends and receives some of its best individuals. The number of migrating individuals depends on the size of the population.

There are two conditions of ceasing the algorithm in the implementation. The program ceases to work if the fixed number of iterations of the main algorithm loop is exceeded. The second condition checks if in a few hundred generations the improvement of the results is relevant enough. If not, the algorithm states that the optimal individual was created and ceases to work.

IMPLEMENTATION INTO GPU PROCESSOR

In order to implement the algorithm into the GPU processor the Microsoft Visual Studio 2008 was used together with NVIDIA CUDA Toolkit 4.0. During the implementation of the algorithm in the CUDA C language (consequently, into the graphic card with its architecture) the algorithm had to undergo some changes. To work efficiently on the GPU processor the program must have high level of parallelity. According to the golden rule by CUDA programmers, there should be 24 times more plots than cores in the graphic card to use the graphic card efficiently. The algorithm was basically written for the NVIDIA GeForce GTX 285. graphic card (it has 240 cores). Hence, the algorithm was transformed to use min. 5760 plots [13], [19].

In comparison to the version for the CPU processor there was one significant change: each individual is served by another thread. The population of 100 individuals results in just 100 threads, which is far too small for CUDA technology. Hence, to increase the number of threads the additional changes were introduced. First of all, the number of individuals reached 5760. However, the tests the CPU version of the algorithm had undergone proved that increasing the number of islands improves the results more than increasing the number of individuals. In order to divide the 'genetic material' between the islands we used the standard CUDA technology mechanism splitting threads into blocks. 60 threads (100 individuals in each) gives 6000 threads which is suitable for CUDA technology for GTX 285 graphic card [6], [18]. To sum up, the entire population was divided into M islands each having N individuals. Each individual is served by the separate thread. Blocks communicate one another by virtue of global memory of the graphic card. Inside the blocks the population is stored in shared memory which is a little bit faster.

The code dedicated to the graphic card is encrypted in kernels which are separate functions. They are executed fully in the GPU processor. Kernels can use only the data kept in the graphic card memory. That memory is allocated by means of

cudaMalloc function. cudaMemcpy() function is responsible for transferring data between RAM and graphic card memory (and vice versa).

TESTS

Test implementation dedicated to the CPU processor was carried out on the processor Intel Core 2 Duo E8400 (3600 Mhz) provided with double physical cores and 6MB second level cache.

The implementation dedicated to the GPU processor was carried out on NVIDIA GeForce GTX 285 graphic card provided with 240 cores and 1024 MB of GDDR3 memory (702 Mhz). To ensure that the measurements are credible in both implementations time was measured in milliseconds.

The implementation dedicated to the CPU processor was run on one thread working on one physical core and on two threads working on two physical cores. In the latter case (2 cores) the population was divided into two islands. Measurement results for the 10,100, 1000 and 10000 individuals are presented below in the Figures: 2,3,4,5.

Each measurement was repeated ten times. The mean of these ten measurements was taken into account (fig. 6). Regarding the implementation dedicated to the CPU processor, to measure the time we use the QueryPerformanceCounter() function. In the version dedicated to the GPU processor the cudaEventRecord() function was used. The length of chromosomes (number of the cities) was the same for both CPU and GPU experiments (excluding the minimal case of 10 cities) and was equal to 100.

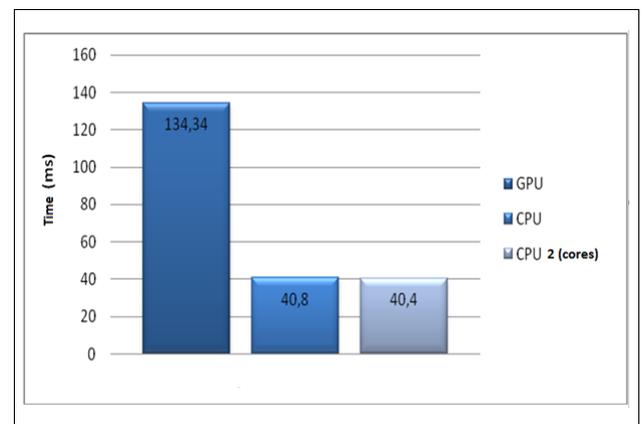


Fig. 2. Measurement results for: the length of the chromosome 10; size of the population 10; number of islands 1; number of generations 100

The time the CPU processor needs for computations raise almost linearly when the amount of data increases but when the graphic card is used it hardly changes. It proves that the graphic card needs a vast number of threads to reach its full efficiency. The GPU processor is not efficient when the number of threads is small because the majority of the cores is unused.

For the sake of the tests the set of points was created consisting of 1024 cities. Each city (point) is given the number from 0 to 1023 as well as cartesian system of

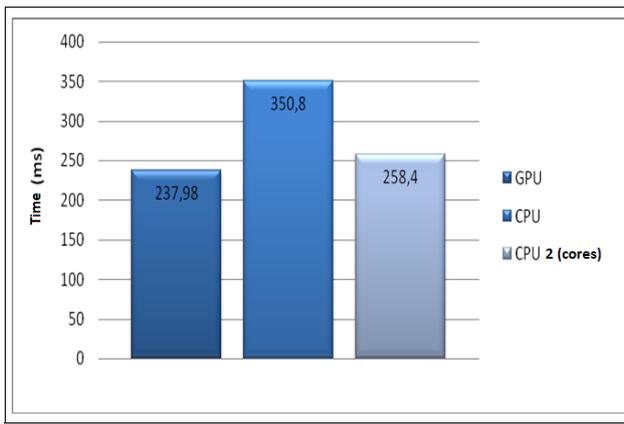


Fig. 3. Measurement results for: the length of the chromosome 100; size of the population 100; number of islands 1; number of generations 100

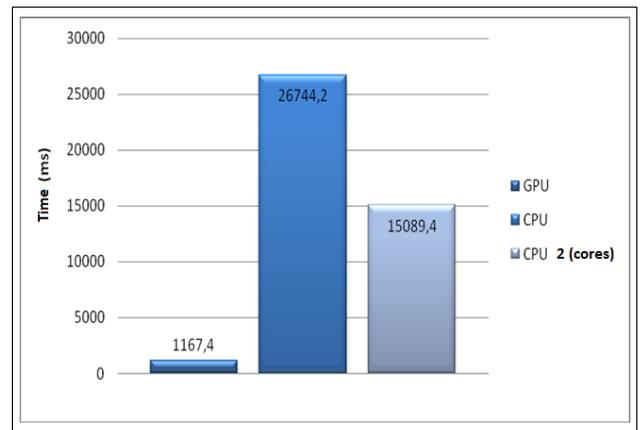


Fig. 5. Measurement results for: the length of the chromosome 100; size of the population 1000; number of islands 100; number of generations 100

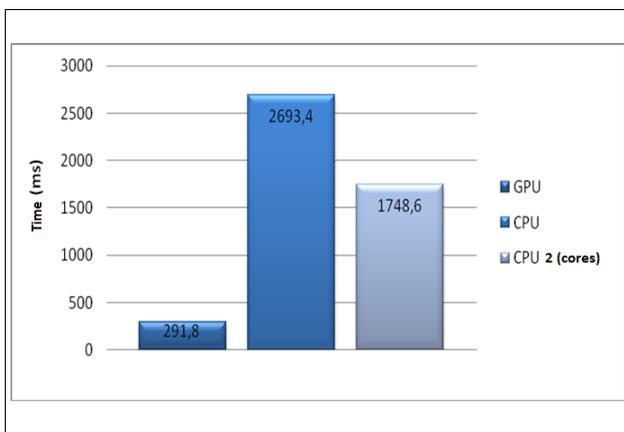


Fig. 4. Measurement results for: the length of the chromosome 100; size of the population 1000; number of islands 10; number of generations 100

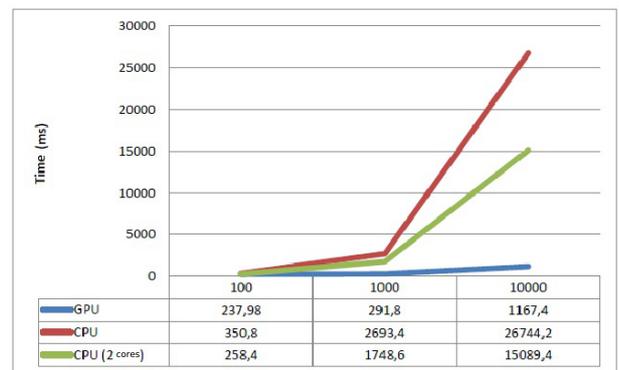


Fig. 6. Time algorithms depending on the size of the population

coordinates x and y. Depending on the chromosome length, always the first N points is used where N stands for the chromosome length. Additionally, in order to enable for the comparison of the effects of the working algorithm, three sets of points taken from TSPLIB library were added. The following sets were chosen berlin52 (52 points), kroA100 (100 points) and kroA150 (150 points)[17].

According to the tests we carried out, graphic cards can be a very good alternative to the standard CPU processors but the problems to solve should be possible to parallel. Moreover, at the small number of threads the efficiency of the both above-mentioned methods are barely distinguishable. Graphic cards require a few dozen of threads to be fully efficient. Using ten thousand of threads enabled us to achieve the acceleration 23-times improved in relation to one CPU core or 13-times improved in relation to two CPU cores.

SUMMARY

The principles of the project were realized. Two versions of the genetic algorithm were created. The first was dedicated to the CPU processor and the second to the graphical processor

GPU. The algorithm is able to solve the travelling salesman problem even for 200 cities which is equal to about $7,89 \times 10^{374}$ possible combinations of the values of genes. The version dedicated to the CPU processor was made parallel by virtue of the OpenMP library. The implementation dedicated to the graphic card was made by means of the CUDA technology. We succeeded in achieving the acceleration 23-times improved in relation to one CPU core and the acceleration 13-times improved in relation to two CPU physical cores. It proves the dormant potential of the contemporary graphic cards. For the parallel computations they became the significant alternative for the traditional CPU processors. Hence, the CUDA technology is more and more often used in supercomputers.

REFERENCES

- [1] Soushil L. J. and Gong L., „Augmenting Genetic Algorithms with Memory to Solve Travelling Salesman Problems”, *University of Nevada, Reno*,2007.
- [2] Holland J. H., „Adaptation in Natural and Artificial Systems”, *MA:MIT Press, Cambridge*,1992.
- [3] Mitchell M., „An introduction to Genetic Algorithms”, *A Bradford Book, Londyn*,1999.
- [4] Goldberg D. E., „Genetic algorithms and their applications”, *WNT*,2003.
- [5] Razali M. N. and Haraghty J., „Genetic Algorithm Performance with Different Selection Strategies in Solving TSP”, *WCE*,2011.
- [6] Sanders J. and Kandrot E., „CUDA example”, *Helion*,2012.

- [7] Goldberg D. E., „Genetic Algorithms in Search, Optimization and Machine Learning”, *Addison-Wesley Longman Publishing Co.*,1989.
- [8] Lawrence D., „Handbook of genetic algorithm”, *Van Nostrand Reinhold*,1991.
- [9] Kim K. and Man F. and Tang and K.S. Kwong and S., „GENETIC ALGORITHMS.: Concepts and Designs Avec disquette Advanced Textbooks in Control and Signal Processing Series”, *Springer*,1999.
- [10] Sivanandam S.N. and Deepa S.N., „Introduction to Genetic Algorithms”, *Springer-Verlag Berlin Heidelberg*, 2007.
- [11] Coley David A., „An introduction to genetic algorithms for scientists and engineers”, *World Scientific*,1999.
- [12] Michalewicz Z., „Genetic Algorithms + Data Structures = Evolution Programs. Artificial Intelligence Series”, *Springer-Verlag*,1992.
- [13] NVIDIA Corporation, „cuda toolkit documentation”, <http://developer.download.nvidia.com>,2011.
- [14] Camilo Rostoke, „Travelling Salesman Problems”,<http://top500.org/list/2011/11/100>,2012.
- [15] Camilo Rostoke, „Travelling Salesman Problems”,<http://www.cs.ubc.ca/labs/beta/Courses/CPSC532D-05/Slides/tsp-camilo.pdf>,2012.
- [16] GPGPU, „General-Purpose Computation on Graphics Hardware”,<http://gpgpu.org/developer>,2012.
- [17] Open MP.org, „The OpenMP API specification for parallel programming”,<http://openmp.org/wp/openmp-specifications/>,2012.
- [18] Marek Obitko, „Introduction to Genetic Algorithms”,<http://www.obitko.com/tutorials/genetic-algorithms/crossover-mutation.php>,2012.
- [19] Konstantin Boukreev, „Genetic Algorithms and the Travelling Salesman Problem”,<http://www.codeproject.com/Articles/1403/Genetic-Algorithms-and-the-Travelling-Salesman-Prob>,2012.
- [20] SENGOKU, H. YOSHIHARA, Ikuo, „A Fast TSP Solver Using GA on JAVA”,<http://www.cs.us.es/cursos/ia1-2006/trabajos/arob98.pdf>,2012.
- [21] Brainz.org, „15 Real-World Uses of Genetic Algorithm”,<http://brainz.org/15-real-world-applications-genetic-algorithms>,2012.
- [22] forums nvidia com, „forums nvidia com”,<http://forums.nvidia.com/index.php?showtopic=195749>,2012.
- [23] cs helsinki fi, „cs helsinki fi bresenh”,<http://www.cs.helsinki.fi/group/goa/mallinnus/lines/bresenh.html>,2012.
- [24] econ iastate edu, „econ iastate edu”,<http://www2.econ.iastate.edu/tesfatsi.html>,2012.

AUTHOR BIOGRAPHIES

ANNA PLICHTA She studied comparative literature at the Jagiellonian University and obtained her degree in 2007. She also studied computer science at Cracow University of Technology and obtained her degree in 2010. Currently, she works as a teaching fellow at Cracow University of Technology. Her e-mail address is: aplichta@pk.edu.pl

TOMASZ GAĆIARZ was born in Olkusz. He studied computer science at the AGH University of Science and Technology. and obtained her degree in 1994. Currently, he works as a teaching fellow at the Cracow University of Technology. His e-mail address is: tga@pk.edu.pl

BARTOSZ BARANOWSKI I was born in Cracow. He studied computer science at the Cracow University of Technology and obtained his degree in 2010. Currently, he works at the IT company. His e-mail address is: bbaranowski@onet.pl

SZYMON SZOMIŃSKI He studied computer science at the Cracow University of Technology and obtained his degree in 2010. Currently, he study computer science at the AGH University of Science and Technology. His e-mail address is: szsz@agh.edu.pl

WORKLOAD CHARACTERIZATION OF MULTITHREADED APPLICATIONS ON MULTICORE ARCHITECTURES

Davide Cerotti
DEIB

Politecnico di Milano
via Ponzio 51
20133, Milano, Italy
davide.cerotti@polimi.it

Marco Gribaudo
DEIB

Politecnico di Milano
via Ponzio 51
20133, Milano, Italy
marco.gribaudo@polimi.it

Mauro Iacono
DSP

Seconda Università di Napoli
viale Ellittico 31
81100 Caserta, Italy
mauro.iacono@unina2.it

Pietro Piazzolla
DEIB

Politecnico di Milano
via Ponzio 51
20133, Milano, Italy
pietro.piazzolla@polimi.it

KEYWORDS

multicore; multithreading; performance modeling

ABSTRACT

Multicore architectures are now available for a wide range of high performance applications, ranging from embedded systems to large scale servers deployed in cloud environments. Multicore architectures are usually subject to two conflicting goals: obtaining a full utilization of the cores while achieving given performance objectives, such as throughput, response time or reduced energy consumption. Moreover, there is a strong interdependence between the software characteristics of the applications, and the underlying CPU architecture. In this scenario, simulation and analytical techniques can provide solid tools to properly design the considered class of systems: however, properly characterize the workload on multithreaded application in multicore environment is not an easy task, and thus is an hot research topic. In this paper we present several models, of increasing complexity, that can characterize multithreaded applications running on multicore architectures.

I. INTRODUCTION

Since the commercialization of the first microprocessor, the advancement of integration techniques allowed the implementation of a higher and higher transistor density on silicon dies. The increasing abundance of transistors on a single chip enabled the consumer market to adopt advanced architectural solutions, that progressively exploited more stages in pipelines, more functional units per stage, more control units per CPU and finally more cores per die. The availability of low cost multicore CPUs has been a key factor in shaping modern massively distributed computing systems, as the use of multicore CPUs allows: power consumption optimization, heating reduction, parallelism flexibility, process granularity exploitation, spatial packing enhancement. Virtualization technologies finally opened the way to the implementation of cheap, massively shared and virtualized computing facilities,

obtained by means of (basically) off-the-shelf components that can easily and affordably be replaced in case of damages.

While multicore architectures and multithreading introduce significative advantages in terms of potential overall performances enhancements, they also introduce an additional complexity level that should be taken in account when designing performance models. The effects are specially relevant when dealing with critical systems or real time applications, but they also have to be considered whenever their impact is multiplied by the existence of a big number of instances that are active in a system at the same time, as in the case of massively distributed architectures or Big Data infrastructures.

In this paper the effects of multicore and multithreading are modeled to explore their relevance with respect to their impact as components of more complex architectures. The study is performed by means of analytical and simulative techniques.

The paper is structured as follows: the next section presents the general approach, then the subsequent section gives a glance of the state of the art; two more sections consider the modeling problem by the analytical and simulative point of view, by introducing new complexity elements in the model; conclusions end the paper.

II. MOTIVATION

Predictive performance modeling is a valuable support in the design and maintenance processes of computer based systems. Traditionally, many analytical and simulative techniques are applied, such as Petri nets or queuing networks, or event-based simulation.

An increase in complexity of the systems to be evaluated translates into an increase in complexity of the corresponding models. As far as the structure of a system keeps simple, or at least modular, analytical or simulative techniques can scale and provide models that can be effectively used; but after above a specific level of complexity, it is necessary to introduce approximations into models, to bypass the natural limits of the chosen technique (or, obviously, to switch to another type of analysis). This, however, can lead to less detailed system

descriptions, and generally requires a complete redesign of the models. Of course approximations have to be carefully studied and justified, as their effects must not destroy the trustfulness of the model.

A possible approach to introduce approximations is to separately evaluate the effects of different layers or components of the architecture, to characterize their performances and find out if they can be somehow neglected or represented with simplified models with respect of the overall architecture.

In this sense, the goal of this paper is to support a wider modeling technique that aims to characterize generalized performance metrics of multicore CPU with a limited complexity. In this way, we can target massively distributed computing architectures, devoted to cloud and Big Data applications, designed to enable the description and the analysis of systems composed by a large number of independent components at different scale (see [1] [2] [3]). In this paper we separately explore the effects of multicore CPU and multithreaded applications, to understand if and how they should be significantly considered in modeling the target architectures.

In particular, we start by proposing characterization that exploits only exponential transition, for which both analytical and simulative techniques can be applied. Then we consider more complex models, which use fork and join of jobs and non-exponential durations, to better capture the relations between cores, threads and number of processes in parallel execution. Such models however, can only be solved via simulation, since analytical techniques would require a number of states that is too large to be considered.

III. RELATED WORKS

The problem represented by performance issues in systems based on multicore CPUs has been analyzed in literature by different points of view. To improve the performance, commercial architectural solutions have been designed and implemented like the Intel Multicore [4] or the AMD 16H [5] architectures. Moreover, several prototypes have been proposed and evaluated such as a utility-based mechanism that partitions a shared cache between multiple applications [6], or a cache-integrated network interface suitable for scalable multicores [7].

Even a proper evaluation of the benefits of these prototypes may be a challenging task due to the complex interaction between several factors. In the most of the literature instead of providing a complete characterization, just effects of one single component of the CPUs (or the system) have been investigated. For instance, [8] shows the different performances achieved when implicit or explicit cache management was used, whereas the authors in [8] propose a multiple linear regression model to investigate the impact of simultaneous multithreading on the performance. Also the effects of internal scheduling in memory was considered in [9] and [10]. Furthermore, the introduction of virtualization techniques, which can be modeled for instance by queuing networks as done in [11] and [12], can have a significant impact on the CPU performance [13]. As stated in [14] several factors introduced

by virtualization affect the performance in way that are still not well-understood.

As the general approach is founded onto the definition or the application of benchmarks that are run on real systems to tune analytical or simulative models, in this paper in vivo measures will be used to validate the proposed models, to obtain a reliable base on which more general performance consideration can be carry out (as, e. g., in [15]), and try and get some general indications about the influence of multithreading and multicore on the overall performances of a complex system architecture.

IV. ANALYTIC RESULTS

We start by focusing on single threaded applications running in multicore environments. A classical queuing model for a multi-processor system running a closed workload, with N identical jobs that are characterized by exponentially distributed service times from CPU and I/O operations is shown in Fig. 1. In particular, the system can be modeled by two stations: one single server that represents the I/O component, and a multiple server that represents the CPU and the scheduler of the operating system. The multiplicity of the server of the queue corresponds to the number of cores of the CPU. As known, the behavior of the system strongly depends on

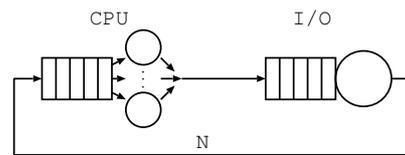


Fig. 1. CPU-I/O Queuing Model.

the type of the workload, i.e.: I/O-bound or CPU-bound. When considering multicore CPU, the influence become even stronger. To properly study this effect, we start by focusing on applications having a total execution time T . We measure how much an application is I/O-bound as the fraction of time α the program spends doing I/O. In this way, if $\alpha = 0$, the workload is completely CPU-bound, and if $\alpha = 1$ the program does only I/O. In particular, we set in the queuing network of Fig. 1, the demands¹ of the I/O ($D_{I/O}$) and CPU (D_{CPU}) stations to:

$$D_{I/O} = \alpha T, \quad D_{CPU} = (1 - \alpha)T. \quad (1)$$

In Fig. 2a, the average system response time, as function of the number of jobs in the system N , is shown for a CPU-only application ($\alpha = 0$), running on a processor with 1, 2, 3 and 4 cores. As it can be seen, since the considered job is entirely CPU-bound, the mean response with c core follows a very simple relation that can be obtained in closed form by applying the Little's law:

$$R(N, c, T) = \begin{cases} T & \text{if } N \leq c \\ \frac{T \cdot N}{c} & \text{if } N > c \end{cases} \quad (2)$$

¹With the term *demand* we characterize the total time spent by an application in one of its phases

The behavior has a very simple explanation: if $N \leq c$, there is at least a core for each job, so no queue is ever formed, and the response time of each job is equal to its demand. If $N > c$, the available cores need to be shared among the considered jobs. The total available capacity of the c cores is thus uniformly distributed among the jobs, leading to a response time $R = T \times N/c$. When I/O is present, ($\alpha > 0$), the CTMC underlying the model shown in Fig. 1 is isomorphic to the one corresponding to a M/M/c/N, where the arrival rate is equal to the inverse of the I/O demand, and the service rate corresponds to the inverse of the CPU demand. In particular, if we define $\lambda = \frac{1}{D_{I/O}}$, $\mu = \frac{1}{D_{CPU}}$, and we call π_0 the probability of having all the jobs doing I/O and π_N the probability of having all the jobs in the CPU, the mean response time can be computed as:

$$\pi_0 = \left(\sum_{k=0}^N \frac{\lambda^k}{\prod_{i=1}^k (\min(i, c)\mu)} \right)^{-1} \quad (3)$$

$$\pi_N = \frac{\lambda^N}{\prod_{i=1}^N (\min(i, c)\mu)} \quad (4)$$

$$X = \mu(1 - \pi_N) \quad (5)$$

$$R = N/X \quad (6)$$

Eq. 6 is the Little's law, Eq. 5 is the Utilization law applied to the I/O (since $1 - \pi_N$ is the utilization of the I/O), and the last two equations are the queuing lengths probability of a M/M/c/N queue (see for example [16]). When the percentage of I/O starts to increase (Fig. 2b, c and d), the effect of the multicore becomes less and less evident. In particular, when $\alpha = 0.24$, there is no more difference between using a 3 or a 4 cores CPU; for $\alpha = 0.36$, all the systems performs as dual-core; and for $\alpha > 0.48$, there seems to be no more advantage in using more than one core. The reason is that the bottleneck switches from the CPU to the I/O, thus canceling the effects improvements that the increased number of cores can bring to the main processor unit. This type of bottleneck switch is further emphasized in Fig. 3, where the effect of parameter α on response time is shown for a workload $N = 20$ when considering $c = 1 \dots 4$ cores. In particular, when the system is single core, the behavior is perfectly symmetric, since the bottleneck switches exactly at $\alpha = 0.5$. Instead when the number of cores increases, the bottleneck switch point tends to reach $\alpha = 0$, that is: even small percentage of I/O can reduce the improvements provided by a higher number of cores. Let us call α^* the value of α at which the bottleneck switches from CPU to I/O. At α^* , both the CPU and the I/O have exactly the same demand:

$$\frac{D_{CPU}}{c} = D_{I/O} \quad (7)$$

By inserting in Equation 7 the definitions of the demands as function of the total service time T and α^* , we obtain:

$$(1 - \alpha^*) \cdot T = c \cdot \alpha^* \cdot T$$

from which we can easily determine α^* :

$$\alpha^* = \frac{1}{c+1} \quad (8)$$

Summarizing, increasing the number of cores the fraction of CPU required for it to be the bottleneck, decreases hyperbolically. For example, with 8 cores, an I/O part that is around 12%, is enough to move the bottleneck away from the CPU, and thus reducing the benefits provided by more cores. Figure 3 shows the location of α^* determined by Equation 8 with a cross placed over the response time curve. As it can be seen, the response time has its minimum in all cases for $\alpha = \alpha^*$: that is the *common saturation point*, the point for which the demands of the I/O and of the CPU are equal. By inverting the definition of α^* , we can determine the best number of cores that can be used with an application, provided that we know its demand in term of CPU and I/O. In particular, the best number of cores c^* can be computed as:

$$c^* = \left\lceil \frac{D_{I/O}}{D_{CPU}} \right\rceil \quad (9)$$

With similar reasoning, we can compute the asymptotic

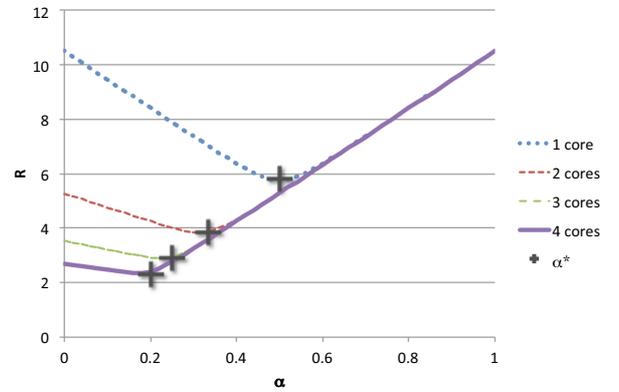


Fig. 3. Response time for $N = 20$ and different numbers of cores as function of α , together with the common saturation point α^* .

bounds R^* to which the system response time tends as function of α and the number of cores c . In particular we have:

$$R^*(N) \geq \max \left(N \cdot \frac{D_{CPU}}{c}, N \cdot D_{I/O} \right) = T \cdot \max \left(N \cdot \frac{1 - \alpha}{c}, N \cdot \alpha \right) \quad (10)$$

V. FITTING CPU AND I/O DEMANDS IN VIRTUALIZED MULTICORE SYSTEMS

To see a practical application of the previous model, we use it to fit the response time of two components of the workload generated by a well known benchmark: the DaCapo suite [17]. In particular we use sunflow and batik: the former is a highly parallel application, which perform ray-trace rendering of 3D computer generated images. The latter produces Scalable Vector Graphics images, using the component Apache Batik: the benchmark is mainly single threaded, even if it can generate several threads during the transcoding of the images elements analyzed by the library. As seen in the analytical results of Section IV, the response time may be mostly influenced by the demand of the CPU or by the demand of the I/O, which in turn depend on the number of cores and

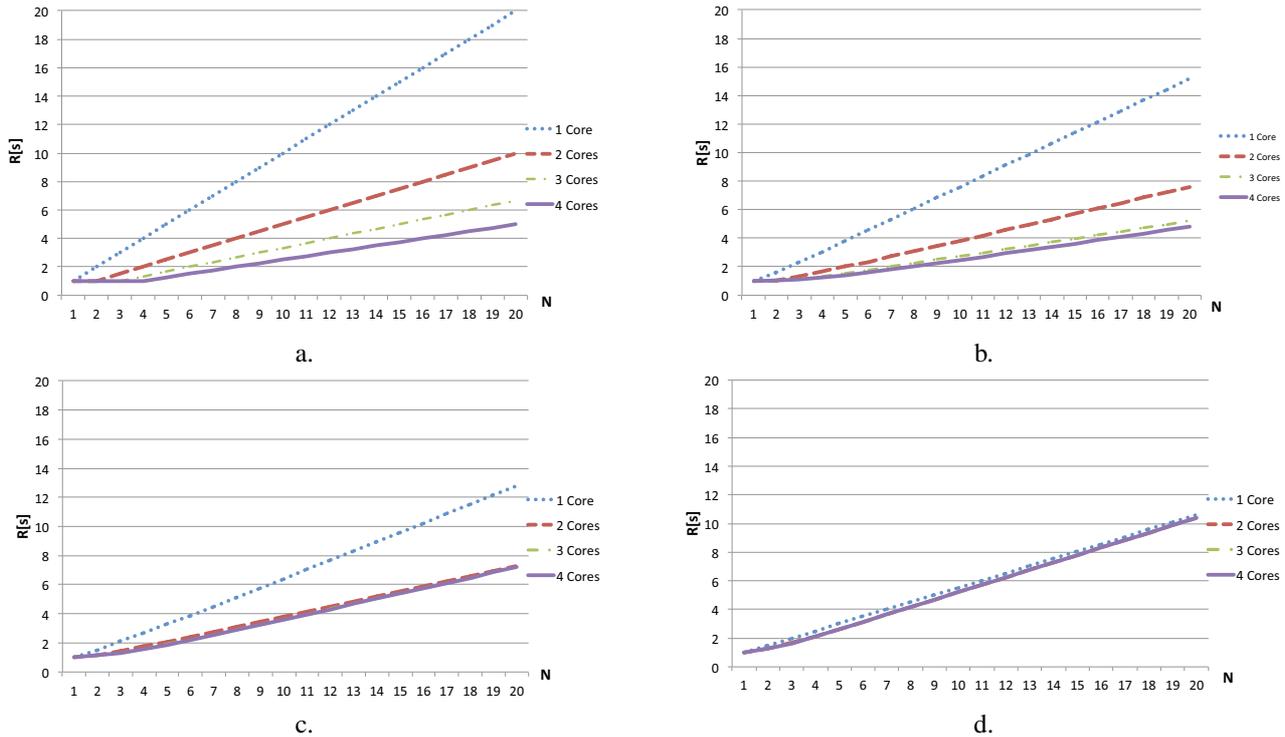


Fig. 2. CPU-I/O switch. a) CPU=100%. b) CPU=76%. c) CPU=64%. d) CPU=52%.

on the fraction of I/O, respectively. We run the benchmarks on Amazon EC2 virtual machines running the Linux OS, and measure the I/O percentage of the applications using the *iostat* command. We then perform a fitting procedure to determine the CPU demand from the collected data using the “GRG solver” in Microsoft Excel, one of the most widely available tools suitable for the considered task. We fit the measures

Simple fit

Benchmark	Mean err.	Demand
sunflow	4,07%	15,524473
batik	17,42%	2,2455743

Load dependent fit

Benchmark	Mean err.	Demand
sunflow	0,85%	15,70436
batik	4,53%	2,2142329

Fig. 4. Fitting errors of the load-independent (upper) and of the load-dependent (lower) models.

of eight-cores VMs against the model presented in Figure 1. We estimate the CPU demands D_{CPU} and $D_{I/O}$ as the values that minimize the squared distance between the model and the measured response times. The fitting results are shown in the left column of Figure 4. As it can be seen, only the fitting with *sunflow* data produces accurate results, with

a mean error of 4%. A visual comparison of the measured response time, and the one obtained with the model is given in Figure 5. The achieved accuracy strongly depends on the characteristics of the benchmark and of the cloud environment: *sunflow* performs 3D rendering, which is a highly parallel and memory consuming task. In this case the L2 shared cache is not exploited at its best, because it becomes saturated almost immediately. Since the M/M/c does not consider caching, the model is capable of describing *sunflow* behavior accurately. *batik* however performs sequential tasks only: this causes an average error of 17.4%, as it can be seen in Figure 6. In order

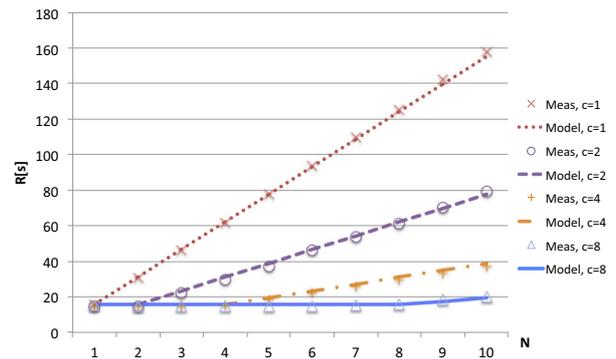


Fig. 5. Comparison of model results and measurements for *sunflow*.

to improve the results, we can use the approach proposed in [18]. In particular, we can consider a load-dependent CPU service station. We call $\lambda(c)$ and $\mu(n, c)$ the I/O and the CPU

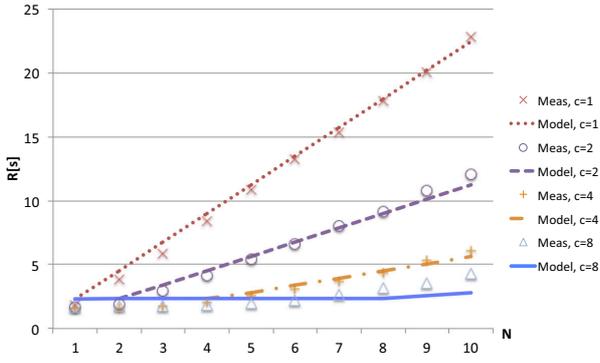


Fig. 6. Comparison of model results and measurements for batik.

service rates when there are c cores available and n jobs in the system. The response time of the proposed model can be computed using Equations 5 and 6, with $\pi_N(c)$ and $\pi_0(c)$ defined as follows:

$$\pi_N(c) = \pi_0(c) \frac{\lambda(c)^N}{\prod_{i=1}^N (\mu(\min(i, c), c))} \quad (11)$$

$$\pi_0(c) = \left(\sum_{k=0}^N \frac{\lambda(c)^k}{\prod_{i=1}^k (\mu(\min(i, c), c))} \right)^{-1} \quad (12)$$

We then repeat the parameter fitting procedure to estimate both $\lambda(c)$ and $\mu(n, c)$. In this way, the model can accommodate for the fluctuations due to the internal architecture of the CPU when the load is lower than the number of cores. The effective service rate remains constant when the load becomes greater than the available cores: from that point on, the CPU is fully utilized, and its behavior operates at a speed which can be considered constant. Results are shown in Figure 7, while errors are reported in the lower part of Figure 4. As it can be seen, the mean error rate reduces significantly for both benchmarks.

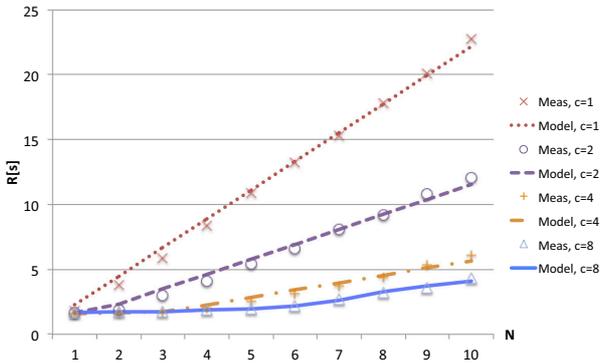


Fig. 7. Comparison of results for the load-dependent model and measurements for batik.

CONSIDERING MULTITHREADED APPLICATIONS

As shown in [19], multicore architectures are particularly efficient when running multithreaded applications: in this case,

when the load is less than the total number of available cores, threads can be run in parallel to reduce the overall response time. For example, in Figure 8, the response time of the *sunflow* benchmark is shown for different number of threads and cores combinations. Since *sunflow* computes rendering of images, it is highly parallelizable: the picture can be segmented into independent parts that can be produced in parallel. As it can be seen, when the computation is split into several threads, the response time is reduced when the product between the number of jobs and the number of threads is less than the available number of cores. As soon as the number of jobs is greater than the number of cores, there is no difference between the response times independently on the number of threads.

Figure 9 shows instead the response time of the *batik* benchmark. Since *batik* is not parallel and it is composed mainly by serial tasks, there are no appreciable differences in the response times with respects to the number of threads. Instead there is a difference when considering a number of cores greater than the total number of jobs.

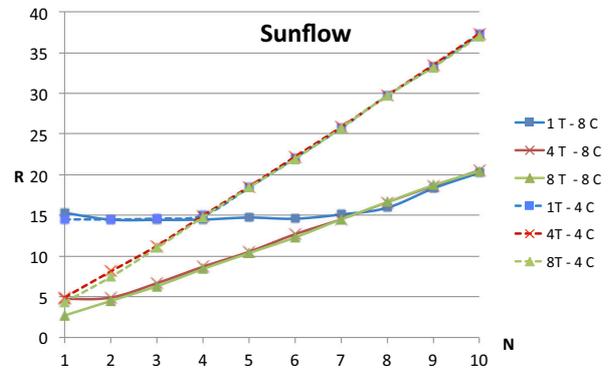


Fig. 8. *sunflow* mean response time in seconds under different configurations of cores/threads, increasing N.

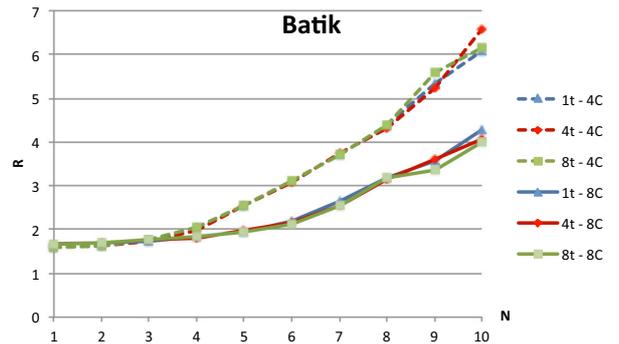


Fig. 9. *batik* mean response time in seconds under different configurations of cores/threads, increasing N.

The model presented in Figure 1 is not able to correctly characterize the parallelism due to the threads. We then propose an extended model, based on the simplifying assumption that a job executes both a serial and a parallel section, as

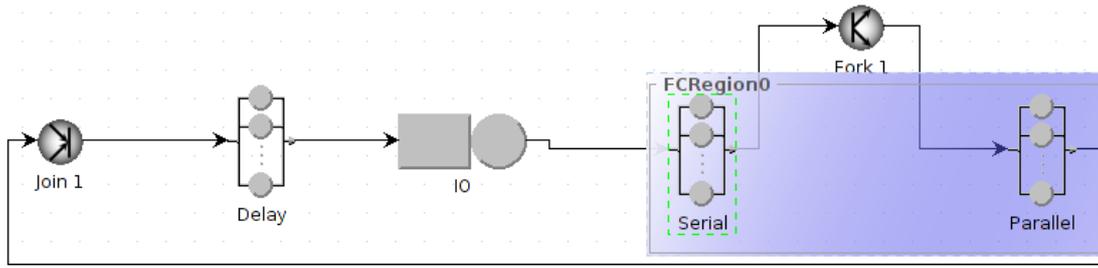


Fig. 10. Model for multithreaded applications in multicore environments.

shown in Figure 10. Each job starts executing the I/O phase, implemented by a standard queue, with a single servant and a First Come First Served discipline. Then, the job executes the serial part, and after that it forks into as many requests as the number of threads that composes the parallel part of the application. Each job is then served by another queue that represents the parallel part of the process. Finally, all the jobs generated by a fork are merged back into a single one using a join operation. The main difficulty is that the service of both the serial and the parallel part is the same: the two queues represents instead the same CPU. This is solved by introducing a *finite capacity region* (FCR), that is a constraint on the total population that is contained in a subset of the model. In particular, we fixed a constraint that tells that the total number of jobs inside the FCS is less or equal to the total number of cores. Then, to allow the parallelism of the cores, both the serial and parallel components of the CPU are modeled using infinite server semantics. Even if the model is very simple, the introduction of the fork and join and of the FCR, prevents it from being solved with analytical and numerical techniques (unless for very small value of the parameters). We thus resort to discrete event simulation: in particular we use the *jSim* component of the JMT - Java Modeling Tool [20]. The considered tool, requires a *reference station* to compute the performance indices: we have chosen to model this by explicitly adding an infinite-server station that represents the start-up of the benchmark (called Delay in Figure 10), characterized by a negligible service time. The model is thus completely characterized by six parameters. In particular, the application is defined through four parameters: the I/O duration, the duration of the serial part, the duration of the parallel part and the number of threads in the parallel part (that is, the number of tasks in which a job is split by the fork node). The number of cores that reflect the hardware architecture is modeled by the size of the FCR. Finally, the number of jobs in the system (for what concerns the serial and the I/O parts) is defined by the initial population of the network.

In this case, also the characterization of the parameters is a challenging task. As for the simpler model, we start by estimating the I/O duration using the I/O utilization values collected by *iostat* command. Then, from the CPU time we estimate the value of the demand of the serial and parallel

part. In this case, we adopt different approaches according to the considered benchmark. Since *sunflow* is highly parallelizable we estimate the demands of the serial and parallel part using Amdahl's Law [21] and the measurements of the system when it is running one job and as many threads as cores. The results of the matching are given in Figure 11 for the four core case, and in Figure 12 for eight cores. The average relative error of the procedure is about 4.47%, similar the one obtained with the simple model of single threaded application considered in Figure 4.

However for *batik* such approach does not provide good results: in this case determining the demands from the measured response time with a single job in the queue can match either the left or the right part of the curve, depending on whether we give more importance to the system with just one thread, or to the system with as many threads as the number of cores. To obtain better results, we put the serial part (since the application is not parallelizable) dependent on the number of jobs in the system. We estimate the value using a basic fitting procedure that perform a gradient descent algorithm with a small number of iteration to cope with the complexity of the solution computed via simulation. Results are shown in Figure 13 for the four core case, and in Figure 14 for eight cores. In this case, the procedure is able to obtain an average error of 9.3%.

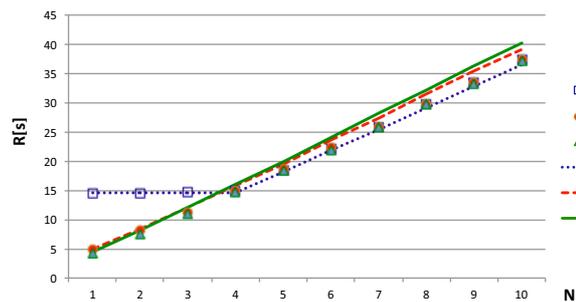


Fig. 11. *sunflow* mean response time in seconds for four cores, and different threads and N .

VI. CONCLUSIONS

In this paper we proposed two approaches to characterize multithreaded applications in multicore environments characterized by a limited number of parameters. Some insights

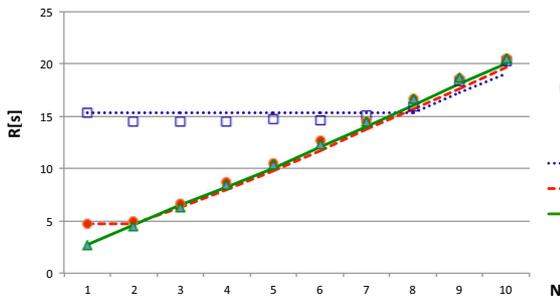


Fig. 12. sunflow mean response time in seconds for eight cores, and different threads and N.

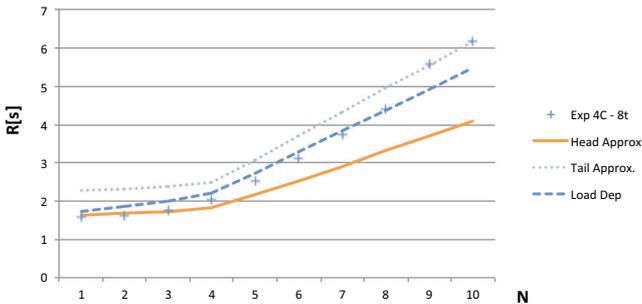


Fig. 13. batik mean response time in seconds for four cores, and different threads and N.

on how such parameters can be derived from measurements coming from executions of real applications on actual multicore machines have been given. The proposed results will be the foundation on which future works targeting Big Data application and cloud infrastructures will be based.

REFERENCES

- [1] A. Castiglione, M. Gribaudo, M. Iacono, and F. Palmieri, "Exploiting mean field analysis to model performances of Big Data architectures," *Future Generation Computer Systems*, no. 0, pp. –, 2013.
- [2] E. Barbierato, M. Gribaudo, and M. Iacono, "Performance evaluation of nosql big-data applications using multi-formalism models," *Future Generation Computer Systems*, vol. to appear, 2013.
- [3] A. Castiglione, M. Gribaudo, M. Iacono, and F. Palmieri, "Modeling performances of concurrent big data applications," *Software: Practice and Experience*, vol. to appear, no. 0, pp. –, 2014.
- [4] J. Doweck, "Microarchitecture and smart memory access," 2006. [Online]. Available: <http://software.intel.com/sites/default/files/m/3/4/d/6/3/18374-sma.pdf>

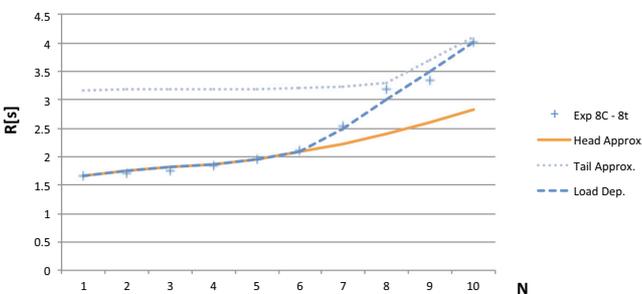


Fig. 14. batik mean response time in seconds for eight cores, and different threads and N.

- [5] "Software optimization guide for amd family 16h processors," 2013. [Online]. Available: http://support.amd.com/TechDocs/52128_16h_Software_Opt_Guide.zip
- [6] M. K. Qureshi and Y. N. Patt, "Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches," in *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO 39. Washington, DC, USA: IEEE Computer Society, 2006, pp. 423–432.
- [7] S. G. Kavadias, M. G. Katevenis, M. Zampetakis, and D. S. Nikolopoulos, "On-chip communication and synchronization mechanisms with cache-integrated network interfaces," in *Proceedings of the 7th ACM international conference on Computing frontiers*, ser. CF '10. New York, NY, USA: ACM, 2010, pp. 217–226.
- [8] S. Schneider, J.-S. Yeom, and D. Nikolopoulos, "Programming multiprocessors with explicitly managed memory hierarchies," *Computer*, vol. 42, no. 12, pp. 28–34, Dec.
- [9] F. Liu, X. Jiang, and Y. Solihin, "Understanding how off-chip memory bandwidth partitioning in chip multiprocessors affects system performance," in *HPCA 2010*, Jan., pp. 1–12.
- [10] Y. Kim, D. Han, O. Mutlu, and M. Harchol-Balter, "Atlas: A scalable and high-performance scheduling algorithm for multiple memory controllers," in *HPCA 2010*, Jan., pp. 1–12.
- [11] D. A. Menascé, "Virtualization: Concepts, applications, and performance modeling," in *Proc. of The Computer Measurement Groups 2005 International Conference*, 2005.
- [12] F. Benevenuto, C. Fernandes, M. Santos, V. A. F. Almeida, J. M. Almeida, G. J. Janakiraman, and J. R. Santos, "Performance models for virtualized applications," in *ISPA Workshops*, ser. Lecture Notes in Computer Science, G. Min, B. D. Martino, L. T. Yang, M. Guo, and G. Rnger, Eds., vol. 4331. Springer, 2006, pp. 427–439.
- [13] L. Cherkasova and R. Gardner, "Measuring cpu overhead for i/o processing in the xen virtual machine monitor," in *Proc. of the USENIX Annual Technical Conference*, ser. ATEC '05. Berkeley, CA, USA: USENIX Association, 2005, pp. 24–24.
- [14] N. Huber, M. Von Quast, F. Brosig, and S. Kounev, "Analysis of the performance-influencing factors of virtualization platforms," in *Proceedings of the 2010 international conference on On the move to meaningful internet systems: Part II*, ser. OTM'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 811–828.
- [15] M. Gribaudo, P. Piazzolla, and G. Serazzi, "Consolidation and replication of vms matching performance objectives," in *Analytical and Stochastic Modeling Techniques and Applications*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7314, pp. 106–120.
- [16] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. New York, NY: John Wiley & Sons, 1976.
- [17] S. M. Blackburn, R. Garner, C. Hoffmann, A. M. Khang, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, M. Hirzel, A. Hosking, M. Jump, H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanović, T. VanDrunen, D. von Dincklage, and B. Wiedermann, "The dacapo benchmarks: Java benchmarking development and analysis," *SIGPLAN Not.*, vol. 41, no. 10, pp. 169–190, Oct. 2006. [Online]. Available: <http://doi.acm.org/10.1145/1167515.1167488>
- [18] D. Cerotti, M. Gribaudo, P. Piazzolla, and G. Serazzi, "Flexible cpu provisioning in clouds: A new source of performance unpredictability," in *QEST*, 2012, pp. 230–237.
- [19] D. Cerotti, P. Piazzolla, M. Gribaudo, and G. Serazzi, "End-to-end performance of multi-core systems in cloud environments," in *EPEW*, 2013, pp. 221–235.
- [20] M. Bertoli, G. Casale, and G. Serazzi, "The jmt simulator for performance evaluation of non-product-form queueing networks," in *Proc. of the 40th Annual Simulation Symposium (ANSS)*, 2007, pp. 3–10.
- [21] G. M. Amdahl, "Validity of the single-processor approach to achieving large scale computing capabilities," in *AFIPS Conference Proceedings*, vol. 30. AFIPS Press, 1967, pp. 483–485.

DYNAMIC FACTORY

New Possibilities for Factory Design Pattern

Dawid R. Ireno
Jagiellonian University,
6 Profesora Stanisława Łojasiewicza Street, Kraków, Poland
alamandra007@gmail.com, dawid.ireno@uj.edu.pl, ireno@ii.uj.edu.pl

KEYWORDS

Software development; design patterns; dynamic factory; dynamic languages; intermediate language; byte code; just in time compilation; runtime environment.

ABSTRACT

Software design patterns have been within developers' realm of influence for several years now. They come from every possible direction, indicating the best courses of action for problem-solving, and are well documented in numerous articles, magazines, and books. Some are corner stones, constituting the foundation of software development. Others are highly evolved complex constructions using other patterns as building blocks to bring about higher quality in more challenging situations. After years of experience in the Information Technology industry, every experienced developer has his own way of perceiving certain design patterns which he has used, and heard, read or talked about. But as the years pass, technology evolves, software design pattern knowledge is still not yet finally distilled, and new design patterns are created. In this article a new design pattern, which the author has called the dynamic factory, is explained. The new type of factory enhances the design pattern possibilities known so far. It creates new object types according to the situation, containing just what is needed, and nothing redundant.

INTRODUCTION

The factory design pattern in software manufacturing is a way to implement the object creation process in a situation where a constructor is not preferred. It is the standard manner of encapsulating logic that lies behind an object's construction. Although this seems somewhat straightforward, it can bring about a great deal of misunderstanding.

SOFTWARE DESIGN PATTERNS

A design pattern is a well-known and technologically independent way to solve a family of problems and is carefully documented, proven to be effective and recommended for use in developed projects.

CONSTRUCTORS

As software trends evolved and procedural languages changed to object ones, there becomes a need to construct objects. The foremost method in which to do this are constructors - factory functions, which reside in a class definition and always return objects of the class they are placed in. Constructor methods may be parameter-free, but may also have various arguments. But one thing is certain; it can never return an object of any derived class. Constructors can only produce instances of the exact class in which they are contained.

Let us suppose there is a class named Shape. The line initializing a new instance and assigning it to a variable would look similar to the following in most of today's known languages:

```
var shape = new Shape(); // Declare the shape
// variable, and initialize it with a new Shape
// class instance.
```

This generally means that an instance is created using default initialization logic. If a developer wants to pass some logic to an initialization block it is still as simple as it seems, assuming the class supports it.

```
var color = Color.Blue;
// Built-in enumeration of colors is used
// to initialize the variable.
// Then the color is passed to the Shape
// constructor.
var shape = new Shape(color);
```

If, on the other hand, a developer wants to create a circle, knowing that circle is a shape in its nature, meaning in the inheritance chain, a special convention is needed.

```
// Enumeration of shape types was
// previously declared in the code.
if (type == ShapeType.Circle)
    var shape = new Circle(color);
else
    ; // TODO: Handle other shape types here.
// Developer cannot write
// var shape = new Shape(type);
// because it might only return Shape, but not
Circle.
```

In [5] one can read about still more dangerous creation scenarios.

When the knowledge for creating an object is spread out across numerous classes, you have creation sprawl: the placement of creational responsibilities in classes that ought not to be playing any role in an object's creation.

Therefore, in the following chapter, it is demonstrated how to manage all types of the aforementioned situations, in a more elegant manner - a way in which to omit logical comparisons and the need to possess knowledge of what the derived class type is. In reality, the only necessity is merely to create a derived class instance while having only some parameterization knowledge. Additionally, all the creation logic will be situated in a single location within the code.

FACTORY

An object factory is the simplest manner of solving the aforementioned problem. As mentioned in [1]

A Factory pattern is one that returns an instance of one of several possible classes depending on the data provided to it.

A factory constructs objects of well-known types. Using the factory, the construction logic mentioned in the previous chapter would be much simpler.

```
var type = ShapeType.Circle;
var color = Color.Blue;
var shape = Shape.Create(type, color);
```

Now in the factory construction method, the logical comparisons are undertaken bearing in mind the need to know what the derived class types are. The factory method code would look similar to the following.

```
public Shape Create(ShapeType type, Color color)
{
    var shape = null;
    if (type == ShapeType.Circle)
        shape = new Circle(color);
    else
        ; // TODO: Handle other shape types here.
    return shape;
}
```

The factory method may be static, but that is not a given. It is usually static if placed in a class of which the method produces instances. If not, the class containing the given factory method usually implements an interface defining the factory method. Despite the applied approach, it is better than in the previous example as the object construction code is encapsulated in a single method body. However, in reality this is the solitary advantage of this approach.

It is worth mentioning is that in a factory design pattern, constructors of types returned by the factory are sometimes intentionally not made publicly available. In this case, the factory method acts as a gateway for creating objects of a

certain type. Construction logic is not divided into several classes and is thus much easier to maintain.

However, when the factory method supports an ever-increasing number of creational options because of growing business requirements, factory methods start producing various, only partially similar object sets, so-called object families.

For example, when not only the color of the shape is important, but also its size, dimensionality, and for 3D shapes the density and friction, there would be a resultant factory method constituting numerous arguments; of which some would be useful for all families while others would be used only for a single family. As a consequence, most arguments would have null value, and few would have a value assigned at the same time.

```
var ct = ShapeType.Circle;
var st = ShapeType.Sphere;
var density = 0,7;
// Variables can be of various types,
var friction = 0,1;
// also floating point numbers.
// Use factory methods to create objects.
var circle = Shape.Create
(ct, color, null, null);
var sphere = Shape.Create
(st, color, density, friction);
```

However, the factory pattern is commonly overused, if not understood correctly, as described in [2].

I've seen numerous systems in which the Factory pattern was overused. For example, if every object in a system is created by using a Factory, instead of direct instantiation (e.g., new StringNode(.)), the system probably has an overabundance of Factories.

In the given example, the actual goal was to have different factories produce 2D and 3D objects. It is possible in this situation to have factory methods with different signatures. In this example a 3D factory will have the same base arguments as a 2D factory, but additionally will add its own arguments.

```
var f2d = new TwoDimFactory();
var f3d = new ThreeDimFactory();
var ct = ShapeType.Circle;
var st = ShapeType.Sphere;
var density = 0,7; var friction = 0,1;
var circle = f2d.Create(ct, color);
var sphere = f3d.Create(st, color, density,
friction);
```

Although the proposed code gives us a straightforward implementation process, 2D and 3D factories' codes become separated. Let us observe that factories have the same arguments in part; with that knowledge in mind, a more appropriate solution may arise.

ABSTRACT FACTORY

The abstract factory patterns come to the rescue here. The difference is that there is an abstract class with the factory

method as yet not implemented, but said to produce some type of objects. Shapes will be the example given in this case. The abstract factory is said to produce families of related objects, and the concrete family creation method is implemented by the concrete factory. In [2] an explanation is provided.

If the creation logic inside a Factory becomes too complex, perhaps due to supporting too many creation options, it may make sense to evolve it into an Abstract Factory. Once that's done, clients can configure a system to use a particular Concrete Factory (i.e., a concrete implementation of an Abstract Factory) or let the system use a default Concrete Factory.

In [6] the authors describe the abstract factory pattern using a straightforward example with two distinct factory methods so as to better understand the difference.

If the abstract factory has two methods CreateProductA and CreateProductB, than one subclass of factory (Factory1) will create ProductA1 and ProductB1, and the other subclass (Factory2) will create ProductA2 and ProductB2 because the factory always produces families of related products.

As the abstract factory defines the factory method signature, all concrete factories must maintain compatibility. In this situation, literals such as connection strings in database development, simple object arrays, key-value dictionaries or dynamic objects are used. Let us demonstrate an example using the dynamic objects approach.

```
var f2d = new TwoDimFactory();
var f3d = new ThreeDimFactory();
var ct = ShapeType.Circle;
// Sample 3D shapes listed in figure 1 below.
var st = ShapeType.Sphere;
var circle = f2d.Create(ct,
    new {Color = color});
var sphere = f3d.Create(st, new {Color = color,
    Density = 0.7, Friction = 0.1});
```

Thus, the given solution has evolved into an abstract factory. Different concrete factories work in rather different ways while maintaining the abstract factory method signature. The natural thing is that the difference lies in the handling of the arguments. If a concrete factory receives an argument that it does not understand, it may neglect it or consider it to be an error, throwing exception, what for user means interrupting program execution. In given example, a 2D factory wishes to have a flat shape and color passed to the factory method. The 3D factory, on the other hand, wants to receive a 3D shape of type, color, density and friction.

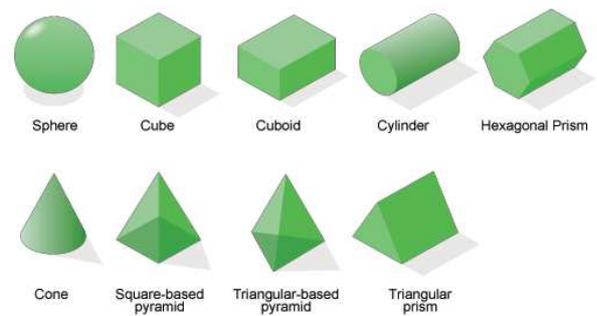


Fig. 1 Sample non-textured 3D shapes

In the example provided, variable information was passed to the factory method using the last argument. This is not obligatory in the case of abstract factories. Usually they have fixed argument vectors and their internal processes are the only factor that differentiates one concrete factory from another, much in the same way that Italian vegetable soup differs from Croatian despite being composed of identical ingredients.

The example given using shapes is extremely simple and does not usually cover the real world system requirements. Therefore one should focus on the 3D factory and assume that he or she wants to change the simple information like shape color to more complex one like shape fixture. To demonstrate the nature of the problem, a fixture will be the complex structure containing information about the texture (image) covering the shape and its luminescence. Let us also assume fixtures are singletons and each fixture points to all shape instances that use this fixture. Although this task seems to be nothing more than that which has been previously mentioned, in the software development industry this topic is covered by a special variation of factory pattern.

COMPLEX FACTORY

Creating objects with an advanced structure is covered by the complex factory design pattern. This type of factory may be abstract, though this is not necessarily the case. Most importantly it creates complex structures of objects, hiding the logic that lies behind object graph initialization. In this section, therefore, the factory function will be described in detail.

```
// Simplified notation
// ReturnedType ClassType.MethodName(Parameters)
// {} is used to denote static methods.
public Shape ThreeDimFactory.Create(type,
    density, friction, texture, luminescence)
{
    var fixture = null;
    // Fixture.All holds all Fixture
    // instances. Check if instance with given
    // parameters was already created or not.
    // If not, create one.
    if (Fixture.All.Contains(texture,
        luminescence))
    {
        fixture = Fixture.All.Get
            (texture, luminescence);
    }
    else
```

```

{
    fixture = new Fixture
                (texture, luminescence);
    Fixture.All.Add(fixture);
}
var shape = null;
if (type == ShapeType.Sphere)
    shape = new Sphere();
else
    ; // TODO: Handle other shape types here.
shape.Density = density;
shape.Friction = friction;
shape.Fixture = fixture;
fixture.Shapes.Add(shape);
return shape;
}

```

Complicated logic has been enclosed here into a single method body for the 3D shape factory. All creations, calculations and object manipulations are done exactly here.

DYNAMIC FACTORY

In this chapter a new design pattern is proposed: The Dynamic Factory. Key ideas about this pattern are explained first. Sample implementation is also provided.

Just in Time compilation

As computer programming languages evolved and virtualized high abstraction environments were created, a need arose to dynamically compile parts of algorithms just before execution. This somewhat lazy method of executable code production was called JIT (Just In-Time) compilation, and also labelled “code jitting” by developers. Various technologies implemented it in different forms. Popular approaches in this area were class and method level compilation types. Among these, the more granular compilation proved to be more useful.

In further considerations, using virtualized runtime environments, so-called virtual machines, will be assumed. This is the key issue while planning dynamic factory implementation in one’s algorithms.

Code templates, static code and runtime types

For further analysis, one has to investigate the purposes for which JIT is utilised. One of many places it has proven to be useful was in template type and method production. When template types or methods are defined, they usually reside in its code base file as parameterized code blocks, useless until the template parameter vector is applied. This static template code cannot be executed earlier than the point at which the type becomes concrete in the virtual machine memory. In such a situation it is desirable to have the code base file as small as possible, but at the same time containing all important information needed for post processing by the Just in Time compilation engine. This way, the static template code is post processed by the Just in Time engine and becomes concrete runtime code in the virtual machine memory. In the case of the template type, it can be instantiated and executed, and is as useful as any other that was non-template type in a code base file. A similar set of circumstances can be viewed in the case of template methods, but on a slightly more granular level.

Dynamic type construction

As Just in Time code compilation proved to be effective, the world became hungry for new methods of runtime type production. This way, dynamic types were created and developers gained the ability to utilize the Just in Time engine to produce type in a way that was previously unknown.

The basic idea is to deliver a way of telling runtime environment to produce a runtime type with given name, which extends the desired base type, implements certain interfaces, and has exactly defined constructors, methods and properties. Although this idea seems rather difficult to cover logically, it turns out the implementation process is not as difficult as had been previously expected.

Dynamic languages

In this article, dynamic languages and interpreters that execute code line by line will not be discussed in any detail, as they are much too slow to meet real business requirements and exhibit poor syntax checking, if there is any at all. These languages are more applicable for dynamic construction types, although their disadvantages place them outside the sphere of author’s interest.

Reflection and emission

When a compiler produces a code base file, it sometimes can prove useful to possess knowledge of how the code works without having the source code itself. This method of browsing outputted files is called reflection, and is usually used for educational purposes.

If a developer does not want anybody to browse his code base files, he uses an obfuscation mechanism to protect them. However, let us assume that the developer will reflect his own files, to learn what assembler instructions and arguments the compiler produces while outputting the code base files. These instructions with arguments are called intermediate code or byte code in environments with a Just in Time-capable virtual machine.

From now on, further investigations are provided using .NET technology and C# language, which is one of many that meet software requirements. Similar implementations can be done in other languages, such as Managed C++, VB.NET, Iron Python, Iron Ruby, Delphi Prism, Oxygene, F-Sharp, J-Sharp or even in other technology such as Java. Single technology is chosen to provide a strict view of the most important factors in implementing a dynamic factory.

It is even suggested that the reader try implementing the pattern on his own, for example in Java. In such a situation, Class Reader, Class Visitor, Annotation Visitor, Field Visitor, Method Visitor, Class Writer, Opcodes and other surrounding built-in types could all prove useful.

To browse non-obfuscated files prepared in .NET technology, one can use Telerik Just Decompile (*), IL Spy (figure 2 below, *), Red Gate Reflector (figure 3 below, *) or Jet Brains Dot Peek.

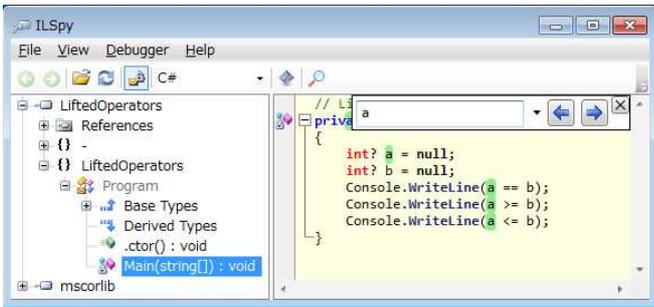


Fig. 2 IL Spy

Tools marked with an asterisk (*) in their current versions have the ability to display intermediate code version of code base files. This is desirable for further investigations. Among interesting tools, Telerik Just Decompile and IL Spy are free decompiling software usable for .NET. Only some versions of Red Gate Reflector are freely available.

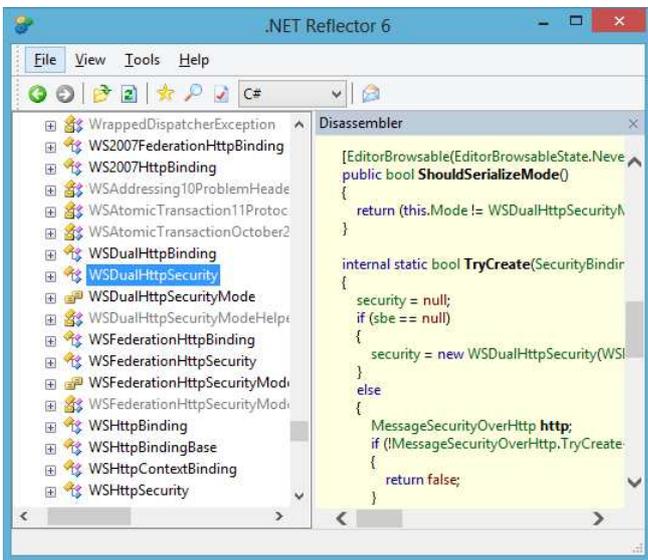


Fig. 3 Red Gate Reflector

An interesting step at this point is to gather knowledge which intermediate code is used to call base methods, pass method arguments, and undertake other low level operations. For sample class with zero-parameter constructors, the code looks relatively simple.

```
namespace SomeNamespace
{
    // Derived type extends base type.
    public class DerivedType : BaseType
    {
        // Derived type constructor calls
        // the base one.
        public DerivedType() : base() { }
    }
}
```

Although using any reflection tool generates intermediate code, it appears to be a bit more complicated than c#.

```
.class public auto ansi beforefieldinit
DerivedType
    extends SomeNamespace.BaseType
{
```

```
.method public hidebysig specialname
rtspecialname
instance void .ctor() cil managed
{
    .maxstack 8
    L_0000: ldarg.0
    L_0001: call instance void
        SomeNamespace.Base::.ctor()
    L_0006: nop
    L_0007: nop
    L_0008: nop
    L_0009: ret
}
```

Code emission is the process of generating intermediate code instruction by instruction, just as the compiler had done for the sample code in C# given previously. Code emission allows for the creation of method bodies, constructors, indexers, properties getter and setter logic, indeed everything supported by the given technology.

Implementing Dynamic Factory

Some approaches to building a Dynamic Factory use built-in dynamic variable type. In .NET the type is called ExpandoObject. The sample factory provided below shows how dynamic objects with tracked properties are constructed.

```
public class Factory
{
    public static dynamic CreateTrackedObject()
    {
        // Class ExpandoObject represents dynamic
        // objects in .NET. It is also referenced
        // using keyword "dynamic".
        dynamic result = new ExpandoObject();
        // Tell the runtime engine to track
        // property changes.
        ((INotifyPropertyChanged)result)
            .PropertyChanged +=
                new PropertyChangedEventHandler
                    (Factory.HandlePropertyChanged);
        // Property change will be intercepted.
        result.Name = "John Smith";
        return result; // Return dynamic object.
    }
    // Method to handle property changes,
    private static void HandlePropertyChanged
        (object sender, PropertyChangedEventArgs e)
    {
        // Write the name of changed property.
        Console.WriteLine("{0} has changed.",
            e.PropertyName);
    }
}
```

Somewhat similar possibilities are delivered for Java Script developers, although in Java Script all objects exhibit similar behavior. As mentioned in previous chapters, there is actually an entire group of languages with dynamic syntax support, or languages that can only interpret code line by line.

Although this construction pattern is relatively simple, in that the developer does not have to know or understand intermediate language, there are nonetheless some major disadvantages. There are issues with speed; constructed

objects are volatile; and they can be broken at any point in the process.

```
dynamic employee = new ExpandoObject();
// Declare and initialize new property.
employee.Name = "John Smith";
// Detach property from its parent object.
((IDictionary<String, Object>)employee)
    .Remove("Name");
```

In the example above, the dynamic object gains a new property, in that it is assigned a value after creation, and finally the dynamic object loses the new property and its value in a single shot. There is no bake method, to say that dynamic object implementation is final, and the object should be made unbreakable from this point.

A much better approach is provided by utilizing code emission, which means producing intermediate code on the fly. In this kind of factory implementation, the developer needs to know what kind of intermediate code the compiler is producing. To gather this knowledge, reflection is used, as explained previously.

In order to demonstrate dynamic factory strength, new runtime type that derives the given base type, has exactly the same public constructors, and appends few desired attributes to the class definition, will be constructed. It is worth mentioning that this is merely an example of the power provided to the developer; pure proof of the concept, although taken from the real-world application.

First of all, the code emission infrastructure must be configured.

```
IEnumerable<CustomAttributeBuilder> cubs = null;
Type baseType = null;
string fullName = null;
// TODO: Fill in custom attribute requirements,
// base type and dynamic type full name from
// factory method arguments.
AssemblyName an = new AssemblyName();
an.Name = "DynamicAssembly";
// Current application domain will load
// new type.
AppDomain ad = Thread.GetDomain();
// Define in-memory dynamic assembly.
AssemblyBuilder ab = ad.DefineDynamicAssembly
    (an, AssemblyBuilderAccess.Run);
ModuleBuilder mb = ab.DefineDynamicModule
    ("DynamicModule");
// Define new type deriving from base one.
TypeBuilder tb = mb.DefineType(fullName,
    TypeAttributes.Public, baseType);
ConstructorInfo[] cis = baseType.GetConstructors
    (BindingFlags.Public | BindingFlags.Instance);
```

Afterwards, the constructors' intermediate code must be emitted. For simplicity it is assumed that derived classes only call base constructors and does nothing more. Knowledge from the chapter on reflection will be utilized herein.

```
// Iterate through base type public
// constructors.
foreach (ConstructorInfo ci in cis)
{
    // Gather constructor argument type
    //collection.
```

```
Type[] constructorArgumentTypes=
    ci.GetParameters().Select
        (pi => pi.ParameterType).ToArray();
// Define public constructor with the same
// arguments.
ConstructorBuilder cb = tb.DefineConstructor
    (MethodAttributes.Public,
        CallingConventions.Standard,
        constructorArgumentTypes);
ILGenerator il = cb.GetILGenerator();
// Emit intermediate language line by line.
il.Emit(OpCodes.Ldarg_0);
int parameters = ci.GetParameters().Count();
// Load constructor arguments onto the stack.
if (parameters >= 1) il.Emit(OpCodes.Ldarg_1);
if (parameters >= 2) il.Emit(OpCodes.Ldarg_2);
if (parameters >= 3) il.Emit(OpCodes.Ldarg_3);
for (byte i = 4; i <= parameters; i++)
    il.Emit(OpCodes.Ldarg_S, i);
// Call the base constructor.
il.Emit(OpCodes.Call, ci);
il.Emit(OpCodes.Nop);
il.Emit(OpCodes.Nop);
il.Emit(OpCodes.Nop);
il.Emit(OpCodes.Ret); // Return the derived
// type instance.
}
```

What remains is to append the desired attributes to derived class definition and bake the new runtime type. This type has all the required features, is extremely quick and non-volatile.

```
if (cubs != null)
    foreach (CustomAttributeBuilder cub in cubs)
        tb.SetCustomAttribute(cub);
derivedType = tb.CreateType();
```

Note that merely compiling the code does not automatically mean that it will work as expected. Dynamic factories should be rigorously tested before use in business environments.

In order to further improve performance, new dynamic types should be cached using concrete vector of parameterization arguments. In the given example it would be the vector <Base Type, Full Name, Type Attributes>. Although it is important to note that two types with the same Full Name of the type cannot exist in one application domain. For reasons of simplicity, the type Full Name will be used as a key in the cache dictionary. The dynamic factory will be able to produce instances after the dynamic type is retrieved from cache or created and baked on the fly. When creating instances, the desired constructor will be automatically best fitted investigating constructors argument types, as if instances of base type were being created.

```
public static Dictionary<string, Type> Cache =
    new Dictionary<string, Type>();
public static T Create<T>( string fullName,
    IEnumerable<CustomAttributeBuilder> cubs,
    params object[] constructorArguments)
{
    Type baseType = typeof(T);
    Type derivedType = null;
    // Check if cache dictionary contains desired
    type.
    if (Cache.ContainsKey(fullName))
        derivedType = Cache[fullName];
```

```

else
{
    // TODO: Use mentioned logic to create new
    // type. New type is baked and therefore
    // non-volatile.
    // Store the new type in cache dictionary.
}
// Create instance using desired constructor.
T result = Activator.CreateInstance
    (derivedType, constructorArguments);
return result;
}

```

This kind of dynamic factory implementation was utilized in two business scenarios by the author. In both cases, modified versions were used so as to fit specific business needs, while maintaining the core principles as discussed. The most interesting case was the need to create transactional objects that implemented certain behaviors. A dynamic factory was used to provide object types with the desired structure and functionalities, and change tracking/recording code injected into the implemented mechanisms of the instances of constructed type. Tracking functionalities were then utilized to implement transactional behavior. While recording changes to the state of the objects, they offered the ability to roll-back all actions up to a certain point in time – namely the moment when all previous transactions have been successfully committed. In both business scenarios mentioned by the author, the dynamic factory proved to be extremely useful.

Inversion of Control

Last but not least, it is worth mentioning what Inversion of Control means. Basically, it is a method of constructing and utilizing objects. In this technique the most important factor is that object coupling is bound at the time of code execution. Using object reference analysis in code, it cannot be predicted which objects will cooperate. In [7] the authors write:

The function of IoC is transferring the control from code to external container. [...] Relationship between the components is specified by the container in the runtime.

An Inversion of Control container is a design pattern used to localize objects that should be used in certain situations. It may be utilized in conjunction with patterns that construct instances of new objects if they do not yet exist in the container.

CONCLUSION

Let us summarize the collected knowledge and think over the final outcome of conducted reasoning.

Constructor

Constructors should be always used whenever possible, assuming that no complicated logic lies behind object construction. Constructors also always return the type they reside in, with no possibility of producing derived type instances.

Factory

A factory is commonly used when there is a need to construct objects according to the environment state. This state is passed on to the factory as a set of variables. Depending on the provided argument values, the factory produces the desired object. Factories can also produce derived type instances. All the construction logic of a factory resides in a single location. This design pattern is commonly known to be overused.

Abstract Factory

This variation of factory proved to be useful when creating object families when objects are similar in some aspects, but differ in others. A situation arises when different factories have factory methods with the same signature, but which work in a different manner. Additional logic for creation is passed on using strings, array, dictionaries and dynamic objects.

Complex Factory

A complex factory is a means of creating object graphs. It covers all constructors and object connections logic, and is used in difficult projects to organize structure and make code easier to maintain and enhance. A complex factory is in some aspects similar to the facade design pattern, which will be described in further detail. However, a complex factory does not have to conceal any disadvantages of code. In [1] we can read

Facade is a way of hiding a complex system inside a simpler interface. [...] This simplification may in some cases reduce the flexibility of the underlying classes, but usually provides all the function needed for all but the most sophisticated users.

Dynamic Factory

This kind of factory is used when a developer does not know exactly what the restrictions for object types or functionalities will be. This knowledge is gathered and utilized at the time of program execution. It is the most advanced factory design pattern variation, which creates new object types with only what is required in a certain situation, and instantiates them on the fly. It may be implemented in two ways. The first is when the factory produces fully dynamic but volatile objects; the second, when the factory produces brand new first-baked and non-volatile type instances. The second one is more difficult for the developer, requiring intermediate language knowledge for coding, and the source code is less readable for humans. Although the constructed objects are without the disadvantages of fully dynamic objects, they have therefore proved to be adequate in business environment edge-cutting software constructions.

Summary

Depending on the situation, each factory variation design pattern is considered useful. The more difficult the situation

which is encountered, the stronger the tools which are used. Among them is the proposed dynamic factory pattern with intermediate code emission which gives us the widest range of possibilities. It is more difficult to implement and technology prerequisites are high, met only by modern languages. However, this is the type of technology which will be used in large solutions in upcoming years, and new projects can benefit from this design pattern.

Using a factory pattern always includes the requirement to access factory methods. They are usually implemented as either static or interface. Sometimes objects containing factory methods are implemented as singletons. Another practiced approach is using an Inversion of Control container with a finder method, used to localize the right factory in a given situation. In this case, the factory is implemented inside the Inversion of Control container, as its creational mechanism. It produces objects in concrete situations using the implemented set of rules, which will be described more precisely in an upcoming chapter. Of course, if chosen, the factory used by Inversion of Control may be any kind of factory described in the article.

APPENDIX

Please note that there exist numerous misunderstandings about dynamic factory in literature. For example [3] describes the Inversion of Control container automatic initialization mechanism, using type attributes to localize types that should be instantiated. Although Inversion of Control container with such a mechanism instantiates new objects, it should not be mistaken for any of kind factory pattern. There might be a factory hidden inside an Inversion of Control container, as mentioned previously. It may have some creation rules for certain desired situations. But that does not make such a factory dynamic in any aspect. In [4] on the other hand, the authors propose a factory that reads static types to be instantiated from code base files indicated in XML files or data bases. It is worth mentioning that simple factory variation proposed by the authors has been used in business products like Microsoft Visual Studio or Microsoft Windows Explorer for many years. Of course, although with a substantially different meaning to that in [3], [4] also has nothing to do with dynamic creating new types.

Builder and Complex Factory design patterns are also commonly confused. The builder pattern mentions nothing about the complexity of objects. The most straightforward example is the String Builder commonly known from languages like C-Sharp, C++, Delphi, Java and Java Script. On the other hand, while hierarchically nesting many builders into others, one can obtain an organized structure for creating complex objects. Enclosing this complex creational structure in one easy-to-handle factory method means creating a Complex Factory. Although creating a complex factory does not mean that numerous builders are required, but rather relates to building a facade for the creation of a complex structure.

Further investigations are planned for new constructions and applications of dynamic factory design pattern. Research

will be also conducted in order to support the dynamic creation of new static types using lambda expressions, and anonymous types, methods and delegates. This will partially alleviate the need to code factory methods in pure intermediate language.

ACKNOWLEDGMENT

Special thanks to my students who directed me to cover factory design patterns in more detail – as such patterns are well documented but also rather superficially understood. Otherwise it may have not been quite so clear that covering the deficiencies in the literature was a worthwhile exercise.

AUTHOR BIOGRAPHIES



DAWID R. IRENO was born in Kraków, Poland and attended the Jagiellonian University, where he majored in computer science and earned his master's degree in 2007. During the following years, he worked on business projects for Roche, Microsoft, Comarch and other large companies in various cities around Poland, utilizing .NET,

ASP.NET, JavaScript, ASP.NET MVC, Ext.NET, PowerShell, SharePoint, SQL, LINQ, WCF, WPF and Silverlight technologies. In 2012 he began his Ph.D. studies at the Jagiellonian University, where he is researching stream databases. His webpage can be found at <http://www.powershell.pl/>.

REFERENCES

- [1] James W Cooper "Java Design Patterns" by Addison-Wesley.
- [2] Joshua Kerievsky "Refactoring To Patterns" by Addison-Wesley.
- [3] Romi Kovacs "Design Patterns: Creating Dynamic Factories in .NET Using Reflection" from MSDN Magazine, March 2003.
- [4] León Welicki, Joseph W. Yoder, Rebecca Wirfs-Brock "The Dynamic Factory Pattern", Proceedings of the 15th Conference on Pattern Languages of Programs, 2008.
- [5] Joshua Kerievsky "Refactoring to Patterns" by Addison-Wesley, 2004.
- [6] A. A. Nykonenko "Using design patterns in computer linguistics: Creational patterns. Part I: Abstract Factory and Builder", Cybernetics and Systems Analysis, Vol. 48, No. 1, January, 2012, pages 138-145, by Springer Science+Business Media, Inc.
- [7] Ke Ju and Jiang Bo 'Applying IoC and AOP to the Architecture of Reflective Middleware', 2007 IFIP International Conference on Network and Parallel Computing - Workshops, pages 903 to 908.

COPYRIGHTS

Fig. 1. http://www.bbc.co.uk/bitesize/ks3/maths/shape_space/3d_shapes/revision/2/

MEMETIC COMPUTING IN SELECTED AGENT-BASED EVOLUTIONARY SYSTEMS

Aleksander Byrski, Marek Kisiel Dorohinicki
Department of Computer Science
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Krakow, Poland
Email: {olekb,doroh}@agh.edu.pl

KEYWORDS

evolutionary algorithms; continuous optimisation; multi-agent computing systems; memetic computation

ABSTRACT

In the paper an application of selected agent-based evolutionary computing models, such as flock-based multi agent system (FLOCK) and evolutionary multi-agent system (EMAS), to the problem of continuous optimisation is presented. It turns out, that hybridizing of agent-based paradigm with evolutionary computation brings a new quality to the meta-heuristic field, easily enhancing static individuals with possibilities of perception and interaction with other agents. The examination of selected benchmarks leads to the observation regarding the overall efficiency of the systems in comparison to the standard genetic algorithm (as defined by Michalewicz) and memetic versions of all the systems. The experiments confirm that the efficiency is dependent on the problem, however, the observed number of fitness function calls makes EMAS dominate over its competitors. This feature makes EMAS a promising solution for the problems with complex fitness functions, (such as inverse problems).

INTRODUCTION

Recently both software agents and evolutionary computation have been gaining more and more applications in various domains. The key concept in multi-agent systems (MAS) constitute intelligent interactions. Evolutionary computation can be perceived as a universal technique for solving optimisation problems. This paper concerns a hybrid evolutionary-agent approach. In contrary to typical approaches reported in literature (see e.g. [17] or [8] for a review) we assume that evolutionary processes are incorporated into a multi-agent system at a population level [10]. The advantages of agents autonomy in this case appear in the possibility of enhancing evolutionary processes with agents interactions, e.g., making possible undertaking autonomous decisions regarding the reproduction by choosing the partner agents.

The paper aims to present selected results of the experiments regarding the selected evolutionary agent-based computing systems. The stress is put on evolutionary multi-agent systems (EMAS), which over the years proved useful in different optimisation problems (e.g., single-criteria, multi-criteria, discrete, continuous) [3].

In this paper, one of the most important features of EMAS is presented—a relatively low computational cost measured as a number of fitness function calls. This makes the system appear well-suited for the problems utilising complex fitness function, requiring e.g., running a simulation to compute the value of the fitness (see inverse problems [1]). This conclusion is based on premise of the presented experimental results concerning popular continuous optimisation benchmarks in comparison to two selected algorithms, popular simple genetic algorithm operating in real-value space [13] and flock-based evolutionary system [11] being another agent-based computational technique proposed by the authors. All the presented algorithms are examined in memetic and standard versions (i.e. with local-search technique enabled or disabled).

In the course of paper, after recalling the basics of evolutionary, memetic and agent-based computation and presenting the concepts of the examined systems, the experimental results are given and discussed, and in the end, the conclusions are drawn.

Agent-based computing paradigm has already been studied, and supported by a number of scientific projects. One of such notable examples is ParaPhrase¹, focusing on supplying hybrid CPU/GPU computing infrastructure via dedicated virtualisation tools. The computing experiments presented in this paper may be treated as preliminary results, planned to be adapted and ported to ParaPhrase infrastructure.

EVOLUTIONARY AND MEMETIC ALGORITHMS

In **evolutionary algorithms** [13] the problem is encoded in a special way (genotype) and random populations of potential solutions are constructed. Based on the existing fitness function (evaluating the genotype), selection is performed (so the mating pool is created) and based on the mating pool, the subsequent population is created with use of predefined variation operators (such as crossover and mutation). The process continues until some stopping condition is reached (e.g., number of generations, lack of changes in the best solution found so far).

It may be seen, that the population of potential (encoded) solutions of a given problem is decomposed into evolutionary islands (there is also a possibility of migration between them) [6]. Such algorithms are usually called “parallel evolutionary

¹<http://paraphrase-ict.eu>

algorithms” (PEA). The most important fact is that the evolutionary algorithm is common to all islands, all operators are applied one by one, during each of generations, to all parts of the population. After meeting some kind of stopping condition, the best solution so far is presented as the optimal one. One of the main drawbacks of such an approach is global (god-like) selection algorithm—possibilities of its de-globalisation will be described later.

Solving optimisation problems with evolutionary algorithms requires that the following must be defined [2]: appropriate encoding of the solutions, crossover and mutation operators appropriate for the encoding, choosing a selection mechanism, and possibly other components of specialized techniques, like configuring topology of islands and migration strategies for the island model of parallel evolutionary algorithms.

Memetic algorithms [14], [12], [15] are population-based techniques that hybridize other meta-heuristics, usually by integrating local search (LS) within the population-based search engine. One of the most important feature of memetic algorithms increased exploitation ability that must be carefully balanced with exploration power of the population heuristics, in order to retain diversity.

In the most cases, two types of memetic systems are defined [15], [12], [16]:

- Baldwinian evolutionary algorithms—in these algorithms the fitness of the individual is evaluated based not only on genotype, but rather on the genotype of one of its potential successors (after, e.g., applying some local-search technique in the course of mutation of the genotype, being a starting point for this local search) —the genotype of the individual remains intact, in the end).
- Lamarckian evolutionary algorithms—in these algorithms, the fitness of the individual is computed after applying local search method to mutate the genotype of the individual (the genotypes is changed, so Lamarckian evolution may be perceived as applying a complex mutation operator).

One of the main advantages of these systems is usually quick attaining of the target optimum, however applying such complex mutation makes the system focused on the exploitation and because of that, additional methods for enhancing the diversity of the population (even such simple, as fitness sharing or crowding [13]) are desired to retain the balance between exploration and exploitation.

Hybridizing memetics with agent-based approaches leads also to the possibility of controlling certain parameters of e.g., memetic-based mutation, adaptation of their value depending on the observation conducted in the environment etc.

INTELLIGENT DECENTRALISATION: FROM INDIVIDUALS TO AGENTS

A **flock-based architecture** may be treated as an extension of the classical island model of evolutionary algorithm (PEA) providing additional level of organisation of the system [11]. Subpopulations on the evolutionary islands (distribution

units) are divided into flocks, where independently conducted processes of evolution are managed by agents (see Fig. 1). It is possible to distinguish two levels of migration:

- exchange of individuals between flocks on one island,
- migration of flocks between islands.

Also merging of flocks containing similar individuals or dividing of flocks with large diversity allows for dynamic changes of population structure to possibly well reflect the problem to be solved.

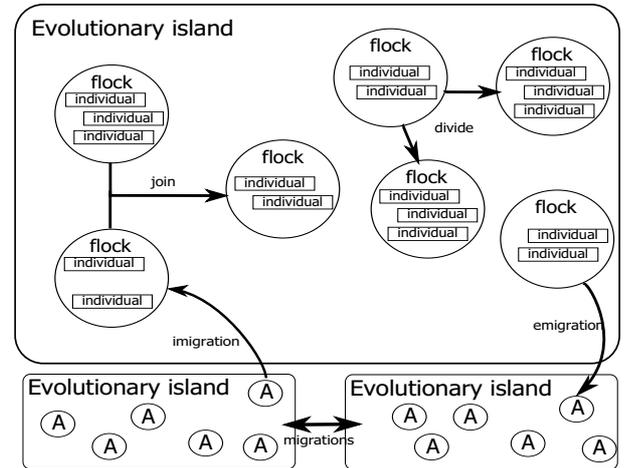


Fig. 1. Flock-based evolutionary system

Agents of an **evolutionary multi-agent system (EMAS)** represent or generate solutions for a given optimisation problem. They are located on islands, which constitute their local environment where direct interactions may take place, and represent a distributed structure of computation (see Fig. 2). Obviously, agents are able to change their location, which allows for diffusion of information and resources all over the system [10].

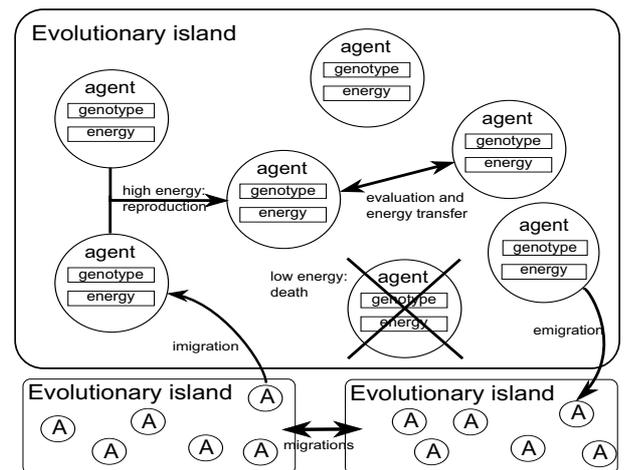


Fig. 2. Evolutionary multi-agent system

Assuming that no global knowledge is available (which makes it impossible to evaluate all individuals at the same

time) and autonomy of the agents (which causes that reproduction is achieved asynchronously), selection is based on the non-renewable resources [7]. Thus a decisive factor of the agent's activity is its fitness, expressed by the amount of non-renewable resource it possesses. The agent gains resources as a reward for 'good' behaviour, and loses resources as a consequence of 'bad' behaviour. Selection is realised in such a way that agents with a lot of resources are more likely to reproduce, while low energy increases the possibility of death.

In the simplest possible model of an evolutionary multi-agent system there is one type of agents and one resource defined. Genotypes of agents represent feasible solutions to the problem.

Energy is exchanged by agents in the process of evaluation. The agent increases its energy when it finds out that one (e.g. randomly chosen) of its neighbours, has lower fitness. In this case, the agent takes part of its neighbour's energy, otherwise, it passes part of its own energy to the evaluated neighbour. The level of life energy triggers actions of death and reproduction (low energy causes death while high energy makes reproduction possible).

Summing up, EMAS agents may perform reproduction action (producing new offspring), death action (in case of low level of energy), evaluation action (in order to exchange the energy based on the fitness function value) and migration action (in order to spread the genetic information among the evolutionary islands). Each action is attempted randomly with certain probability, and it is performed only when their basic preconditions are met (e.g. an agent may attempt to perform the action of reproduction, but it will reproduce only if its energy rises above certain level and it meets an appropriate neighbour).

EXPERIMENTAL RESULTS

In order to examine the features of standard and agent-based computing systems, they were implemented using AgE computing platform (<http://age.iisg.agh.edu.pl>). All parameters of the the systems under consideration (SGA (Michalewicz version [13], FLOCK and EMAS both in standard and memetic versions) were chosen in such way, that the comparison between them could be possible and the perceived differences could depend only on the intrinsic features of the algorithms.

Thus, the configurations of SGA, FLOCK and EMAS were as follows:

- ALL real-value encoding, discrete recombination (offspring gets parents' genes one by one, from each parent with certain probability), normal mutation with standard deviation 0.3 and probability 0.2, stopping condition: reaching 1000th step of the computation.
- SGA 100 individuals, tournament selection.
- FLOCK 5 flocks 20 individuals each, tournament selection, the flocks join together when their populations overlap and divide, when the diversity of the population is low.
- EMAS in the beginning, there are 30 individuals, the population number stabilises at about 100 individuals, starting energy: 30 units, total energy constant:

900 units, reproduction at 15 units, during evaluation agents exchange 5 units, energy of death: 0 units.

- MEM. all memetic versions utilized a Lamarckian mutation based on steepest descent there are three attempts to mutate the genotype, each time the next proposed genotype is sampled three times in the vicinity of the individual, and the best proposition is chosen.

The considered benchmark problems were popular De Jong, Ackley, Rastrigin, Griewank and Rosenbrock functions [9] described in 10 dimensions. All the experiments were repeated 30 times and the standard deviation was computed as a measure of repeatability.

In Fig. 3 the progress of optimisation process conducted in all examined systems was presented. It was displayed as the best fitness observed in subsequent steps of the computation. In order to distinguish individual features of each process, logarithmic scale was used on ordinate axis.

Recalling "no free lunch theorem" [18] the authors were not aiming at proving that one of the examined systems proves as the best for all benchmark used. Instead, certain information about the features of each system may be discovered, when looking at the graphs in Fig. 3 and the tables later on. E.g., quick look at the graphs reveals, that almost independent on the system used, the Rosenbrock problem, being a well known deceptive function, remains the most difficult one. On the other side, De Jong problem, being a simple convex function, appears of course the easiest one to be solved.

When comparing the effectiveness of certain computing systems relatively to the problems solved, looking at the graphs presented in Figs. 3(a), 3(c), 3(e), does not let to favour any of the systems, maybe apart from EMAS doing much better in the case of De Jong function 3(e), though it is to note, that this problem is too straightforward to prove the domination of one of optimisation methods.

When comparing memetic versions of all the examined systems (see Figs. 3(a), 3(c), 3(e)) it is easy to see, that these versions are much better in solving the given problems, than their standard versions, as they reach much better results, moreover, the descent in the direction of the optimum is quicker and the curve depicting it is steeper in the beginning of the computation.

Additional information regarding the efficiency of certain systems may be found in Tables I, II. When looking for the best obtained results throughout the all experiments, it seems, that it is hard to find one algorithm dominating the others (see, [18]).

In Tables III, IV the diversity obtained in 1000th step for the all population was shown. This measure was computed as minimum standard deviation of all genes averaged over the whole population. It is easy to see, that memetic versions of all algorithms tend to process much less diverse populations than their standard versions (a well known problem of memetic computation [15]). Diversity is also quite dependent on the problem, as the problem itself influences the distribution of the populations, see, e.g., column presenting the data gathered for Rosenbrock problem: this values are one of the highest in

the table, as Rosenbrock problem, visualized in 2 dimension as quite flat surface with several bumps, allows the population to be spread more than, e.g., Rastrigin or Griewank problem, where the individuals gather in local extrema throughout the whole computation.

The most interesting results however, are presented in Table V. There, approximated number of fitness function calls computed for all conducted experiments is shown. It is easy to see, that soft selection mechanism (energetic selection) used in EMAS (both in standard and memetic versions) allowed to obtain quite similar results (see, Fig. 3 and Tables I, II) at the same time reducing the number of fitness function calls (better by two-three orders of magnitude when comparing with FLOCK or SGA).

CONCLUSIONS

In the course of the paper selected agent-based computing systems were recalled (FLOCK and EMAS) and the experimental results obtained for optimisation of several benchmark functions were given. Detailed insight into the features presented in graphs depicting the best fitness in the examined population did not allow to state, that one of the tested systems prevailed. However, classical features of memetic computation were spotted: the optimum is pursued faster in the beginning, and the diversity of these systems is lower than in the case of their standard versions.

The most important conclusion is proving, that regardless the efficiency of EMAS in comparison to other systems, it prevails in the means of fitness function calls during the computation (even by two or three orders of magnitude). This feature makes EMAS a reliable means for solving problems with complex fitness functions, such as inverse problems, evolution of neural network parameters (see, e.g., [5], [4]), and others.

In the future the authors plan to enhance the testing conditions by considering continuous and discrete benchmarks as well as increasing the dimensionality of the problems to be solved.

ACKNOWLEDGMENTS

The research presented in the paper was partially supported by the European Commission FP7 through the project Paraphrase: Parallel Patterns for Adaptive Heterogeneous Multicore Systems, under contract no.: 288570 (<http://paraphrase-ict.eu>).

The research presented in this paper received partial financial support from AGH University of Science and Technology statutory project.

REFERENCES

- [1] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. Academic Press, 2005.
- [2] T. Bäck, D. Fogel, and Z. Michalewicz, editors. *Handbook of Evolutionary Computation*. IOP Publishing and Oxford University Press, 1997.
- [3] A. Byrski, R. Dreżewski, L. Siwik, and M. Kisiel-Dorohinicki. Evolutionary multi-agent systems. *The Knowledge Engineering Review*, Accepted for publication, 2012.
- [4] A. Byrski and M. Kisiel-Dorohinicki. Evolving RBF networks in a multi-agent system. *Neural Network World*, 12(5):433–440, 2002.
- [5] A. Byrski, M. Kisiel-Dorohinicki, and E. Nawarecki. Agent-based evolution of neural network architecture. In M. Hamza, editor, *Proc. of the IASTED Int. Symp. on Applied Informatics*. IASTED/ACTA Press, 2002.
- [6] E. Cantú-Paz. A summary of research on parallel genetic algorithms. *IlligAL Report No. 95007*. University of Illinois, 1995.
- [7] K. Cetnarowicz, M. Kisiel-Dorohinicki, and E. Nawarecki. The application of evolution process in multi-agent world (MAW) to the prediction system. In M. Tokoro, editor, *Proc. of the 2nd Int. Conf. on Multi-Agent Systems (ICMAS'96)*. AAAI Press, 1996.
- [8] S.-H. Chen, Y. Kambayashi, and H. Sato. *Multi-Agent Applications with Evolutionary Computation and Biologically Inspired Technologies*. IGI Global, 2011.
- [9] J. Digalakis and K. Margaritis. An experimental study of benchmarking functions for evolutionary algorithms. *International Journal of Computer Mathematics*, 79(4):403–416, April 2002.
- [10] M. Kisiel-Dorohinicki. Agent-based models and platforms for parallel evolutionary algorithms. In M. Bubak, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, editors, *Computational Science – ICCS 2004. Part III*, volume 3038 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2004.
- [11] M. Kisiel-Dorohinicki. Flock-based architecture for distributed evolutionary algorithms. In L. Rutkowski, J. Siekmann, R. Tedeusiewicz, and L. Zadeh, editors, *Artificial Intelligence and Soft Computing – ICAISC 2004*, volume 3070 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2004.
- [12] N. Krasnogor and J. Smith. A tutorial for competent memetic algorithms: Model, taxonomy, and design issues. *IEEE Transactions on Evolutionary Computation*, 9(5):474–488, 2005.
- [13] Z. Michalewicz. *Genetic Algorithms Plus Data Structures Equals Evolution Programs*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1994.
- [14] P. Moscato. Memetic algorithms: a short introduction. In *New ideas in optimization*, pages 219–234, Maidenhead, UK, England, 1999. McGraw-Hill Ltd., UK.
- [15] P. Moscato and C. Cotta. A modern introduction to memetic algorithms. In M. Gendreau and J.-Y. Potvin, editors, *Handbook of Metaheuristics*, volume 146 of *International Series in Operations Research and Management Science*, pages 141–183. Springer, 2 edition, 2010.
- [16] Y.-S. Ong, M.-H. Lim, and X. Che. Memetic computation – past, present & future. *IEEE Computational Intelligence Magazine*, 5(2):24–36, 2010.
- [17] R. Sarker and T. Ray. *Agent-Based Evolutionary Search*. Springer, 2010.
- [18] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

TABLE I. AVERAGE BEST FITNESS OBTAINED IN 1000TH STEP AND STANDARD DEVIATION (COMPUTED FOR 30 RUNS) FOR MEMETIC AND STANDARD VERSIONS OF EXAMINED COMPUTING SYSTEMS (ACKLEY, GRIEWANK AND RASTRIGIN PROBLEMS)

System	Ackley		Griewank		Rastrigin	
SGA Std	0.057 237 12	$\pm 0.007 241 441$	0.085 502 67	$\pm 0.061 758 09$	0.340 349	$\pm 0.107 659 7$
SGA Mem	$2.148 808 \times 10^{-5}$	$\pm 7.754 295 \times 10^{-6}$	0.060 106 21	$\pm 0.029 837 69$	$6.493 436 \times 10^{-8}$	$\pm 4.583 01 \times 10^{-8}$
FLOCK Std	0.076 686 35	$\pm 0.014 271 68$	0.541 716 8	$\pm 0.456 685 6$	1.785 208	$\pm 0.775 546 1$
FLOCK Mem	$4.631 602 \times 10^{-5}$	$\pm 1.580 519 \times 10^{-5}$	0.082 211 74	$\pm 0.058 706 72$	$4.006 674 \times 10^{-7}$	$\pm 2.747 038 \times 10^{-7}$
EMAS Std	0.047 309 22	$\pm 0.148 428 8$	27.977 86	$\pm 18.809 17$	3.011 314	$\pm 1.498 797$
EMAS Mem	0.000 431 029 2	$\pm 0.000 147 903 2$	3.008 084	$\pm 4.320 769$	$3.958 766 \times 10^{-5}$	$\pm 3.085 049 \times 10^{-5}$

TABLE II. AVERAGE BEST FITNESS OBTAINED IN 1000TH STEP AND STANDARD DEVIATION (COMPUTED FOR 30 RUNS) FOR MEMETIC AND STANDARD VERSIONS OF EXAMINED COMPUTING SYSTEMS (ROSENBRACK AND DE JONG PROBLEMS)

System	Rosenbrock		De Jong	
SGA Std	2.706 568	$\pm 1.774 557$	0.001 540 284	$\pm 0.000 387 838$
SGA Mem	2.210 916	$\pm 1.739 984$	$4.583 541 \times 10^{-10}$	$\pm 3.120 165 \times 10^{-10}$
FLOCK Std	3.884 875	$\pm 2.081 934$	0.002 535 208	$\pm 0.000 588 550 1$
FLOCK Mem	0.901 195 7	$\pm 1.036 577$	$1.871 672 \times 10^{-9}$	$\pm 1.202 404 \times 10^{-9}$
EMAS Std	25.200 49	± 101.7666	$1.748 485 \times 10^{-6}$	$\pm 9.186 823 \times 10^{-7}$
EMAS Mem	3.940 108	$\pm 3.054 97$	$6.333 364 \times 10^{-8}$	$\pm 3.935 681 \times 10^{-8}$

TABLE III. AVERAGE DIVERSITY OBTAINED IN 1000TH STEP AND STANDARD DEVIATION (COMPUTED FOR 30 RUNS) FOR MEMETIC AND STANDARD VERSIONS OF EXAMINED COMPUTING SYSTEMS (ACKLEY, GRIEWANK AND RASTRIGIN PROBLEMS)

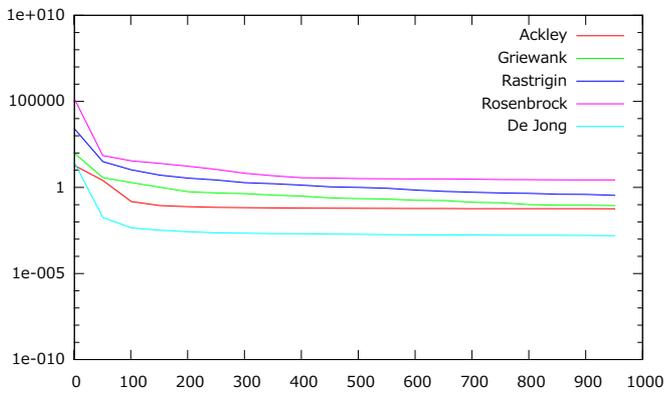
System	Ackley		Griewank		Rastrigin	
SGA Std	0.073 756 12	$\pm 0.017 941 05$	0.073 891 11	$\pm 0.016 700 04$	0.078 326 87	$\pm 0.026 888 11$
SGA Mem	$1.027 91 \times 10^{-22}$	$\pm 2.331 456 \times 10^{-22}$	$6.458 626 \times 10^{-22}$	$\pm 1.118 568 \times 10^{-21}$	$2.841 09 \times 10^{-22}$	$\pm 8.269 988 \times 10^{-22}$
FLOCK Std	0.068 397 06	$\pm 0.012 846 73$	3.282 302	± 3.3432	0.146 949 1	$\pm 0.085 565 59$
FLOCK Mem	$3.269 656 \times 10^{-6}$	$\pm 2.644 61 \times 10^{-6}$	1.244 616	$\pm 2.017 345$	$2.957 484 \times 10^{-6}$	$\pm 3.169 475 \times 10^{-6}$
EMAS Std	0.015 511 87	$\pm 0.006 212 766$	0.098 006 43	$\pm 0.076 427 53$	0.016 464 35	$\pm 0.009 102 824$
EMAS Mem	$2.177 581 \times 10^{-21}$	$\pm 4.648 354 \times 10^{-21}$	0.047 304 02	$\pm 0.065 510 94$	$4.842 205 \times 10^{-21}$	$\pm 1.743 64 \times 10^{-20}$

TABLE IV. AVERAGE DIVERSITY OBTAINED IN 1000TH STEP AND STANDARD DEVIATION (COMPUTED FOR 30 RUNS) FOR MEMETIC AND STANDARD VERSIONS OF EXAMINED COMPUTING SYSTEMS (ROSENBRACK AND DE JONG PROBLEMS)

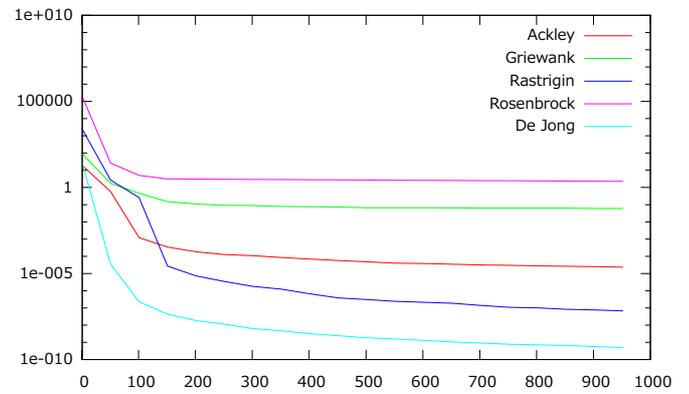
System	Rosenbrock		De Jong	
SGA Std	0.069 791 13	$\pm 0.015 978 85$	0.072 843 56	$\pm 0.013 871 39$
SGA Mem	$7.444 279 \times 10^{-5}$	$\pm 0.000 127 459 5$	$2.007 843 \times 10^{-22}$	$\pm 4.424 505 \times 10^{-22}$
FLOCK Std	0.126 153 6	$\pm 0.088 960 07$	0.071 510 14	$\pm 0.009 508 027$
FLOCK Mem	0.014 950 56	$\pm 0.022 712 15$	$1.288 522 \times 10^{-6}$	$\pm 2.031 521 \times 10^{-6}$
EMAS Std	0.016 752 47	$\pm 0.009 061 646$	0.013 228 51	$\pm 0.006 898 98$
EMAS Mem	$8.764 745 \times 10^{-5}$	$\pm 0.000 389 931 5$	$1.934 059 \times 10^{-21}$	$\pm 3.741 526 \times 10^{-21}$

TABLE V. AVERAGE NUMBER OF FITNESS CALLS DURING 1000 STEPS (COMPUTED FOR 30 RUNS) FOR MEMETIC AND STANDARD VERSIONS OF EXAMINED COMPUTING SYSTEMS

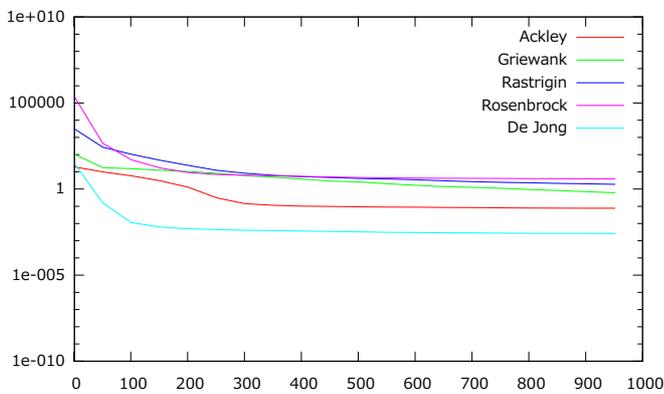
System	Ackley	Griewank	Rastrigin	Rosenbrock	De Jong
SGA Std	100 000	100 000	100 000	100 000	100 000
SGA Mem	299 800	299 800	299 800	299 800	299 800
FLOCK Std	85 386.6664	97 800.0024	93 293.3336	87 506.6668	74 133.3328
FLOCK Mem	255 640	287 200	243 640	211 480	238 559.99
EMAS Std	349.466 87	371.899 98	364.866 71	359.433 51	365.633 22
EMAS Mem	963.001 05	996.299 97	966.602 61	925.199 94	965.600 55



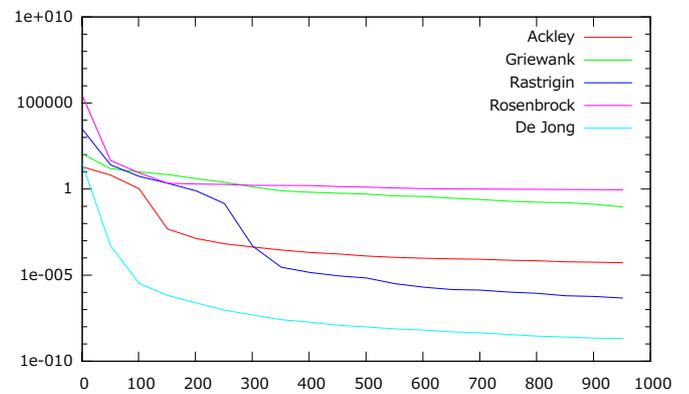
(a) Standard SGA best fitness



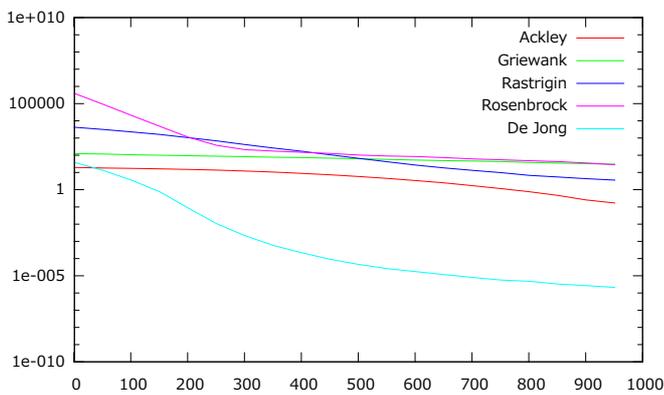
(b) Memetic SGA best fitness



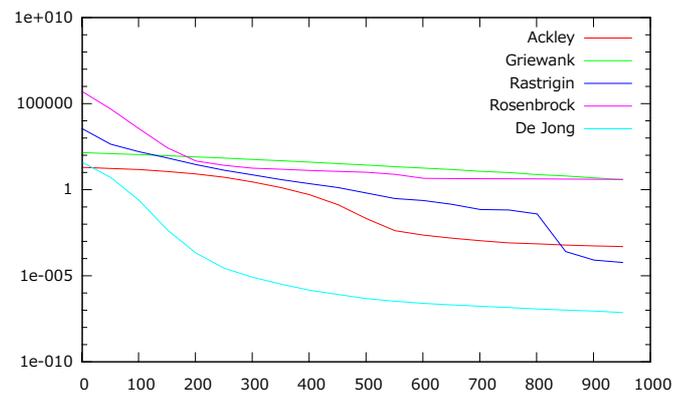
(c) Standard FLOCK best fitness



(d) Memetic FLOCK best fitness



(e) Standard EMAS best fitness



(f) Memetic EMAS best fitness

Fig. 3. Fitness for memetic and standard versions of examined computing systems

OPTIMAL PUMP SCHEDULING BY NLP FOR LARGE SCALE WATER TRANSMISSION SYSTEM

Jacek Błaszczuk*, Krzysztof Malinowski[†]* and Alnoor Allidina[‡]

*Research and Academic Computer Network (NASK)

ul. Wąwozowa 18, 02-796 Warsaw, Poland

Email: Jacek.Blaszczuk@nask.pl

[†]Institute of Control and Computation Engineering

Warsaw University of Technology, Warsaw, Poland

Email: K.Malinowski@ia.pw.edu.pl

[‡]IBI-MAAK Inc., Richmond Hill, Ontario, Canada

Email: Alnoor.Allidina@ibigroup.com

KEYWORDS

Large-scale nonlinear programming; Minimum cost operative planning; Optimal pump scheduling; Water supply; Hydraulic simulations

ABSTRACT

In this paper an operational control for the Toronto's Transmission Water System (TWS) is considered. The main objective of the ongoing Transmission Operations Optimizer (TOO) project consists in developing an advanced optimization and control tool for providing such pumping schedules for 153 pumps, that all quantitative requirements (such as pressure, flow, water level and water quality) with respect to the system operation are met, while the energy costs of delivering fresh water are minimized.

We describe here, in general, the concept of TOO system and the large-scale non-linear, so-called Full Model (FM), based on system of hydraulic equations. The FM model is, in fact, a simplified version of EPANET hydraulic simulator with hourly hydraulic time-step, implemented as a complex NLP optimization model and usually solved over 24-hour horizon to deliver the aggregated optimal solution. To solve the resulting large-scale NLP we use the nonlinear interior-point method implemented in general-purpose large-scale IPOPT solver.

Finally, we included the typical numerical example of application of the TOO Optimizer to solve the 24-hour and 7-day FM problems, and compared obtained optimal FM results with results of hydraulic simulation performed under EPANET simulator.

INTRODUCTION: TOO SYSTEM

The City of Toronto water transmission system is a large complex hydraulic network of treatment plants, pumping stations, storage facilities (including underground reservoirs and elevated tanks) and transmission utilities (including hundreds of kilometers of water mains and significant number of control valves). The City of Toronto water supply system capacity is one of the largest in North America. The Water Supply function is responsible for providing services 24 hours per day, seven days per week. The system consists of treated water pumping at four filtration plants, 29 pumping stations, floating storage at 19 reservoirs and 9 elevated tanks, and approximately 520 km of water mains that transport treated water from the lake up through the system. Water is pumped through a hierarchy of pressure districts with elevated storage facilities. At present a large part of the system within the City of Toronto is essentially manually operated, where an operator decides for example when to turn

a pump on or off. Even when there are no abnormal situations, manual decision making within the City of Toronto system is a complex process.

With this background, the City of Toronto decided to develop the Optimizer that automatically determines control strategies for the Water Transmission System, based on certain criteria, including meeting service delivery levels (pressures, reservoir levels, water quality). The developed TOO Optimizer works as on-line tool alongside the SCADA control systems. The primary objective of the Optimizer (TOO) is to ensure that required water delivery standards are met, while minimizing electrical cost of water pumping. "The aim of pump scheduling is to minimize the marginal cost of supplying water while keeping within the physical and operational constraints, such as maintaining sufficient water within the system's reservoirs, to meet the required time-varying consumer demands." – (Methods, 2003). The complexity of the water system and dynamically changing energy rates present potential opportunities for optimizing operations by minimizing water pumping and treatment costs.

The simplified flowchart diagram of the whole TOO system and its main software element, called the Control Strategy Component (CSC) is presented in figure 1. The CSC component consists of:

- the Optimizer using the FM and FMBM optimization models, and also the library of optimization and matrix solvers
- the EPANET hydraulic simulator using the hydraulic and water quality model of TWS system defined in EPANET INP format
- the FM (NLP) and FMBM (LP) optimization models solved by COIN-OR IPOPT and Clp solvers, respectively
- the library of COIN-OR optimization solvers (IPOPT and Clp) together with the library of HSL matrix solvers (MA27, MA57 and HSL_MA97)
- the Simplifier, which produces a simplified hydraulic model
- the Scheduler generating an optimized pump schedules on the base of FMBM and FM optimal aggregated solutions

The FMBM and FM optimization models are automatically obtained from the hydraulic model of network in EPANET INP format and from additional data (provided by the TOO system) describing operational constraints, electricity tariffs and pumping station configurations. In order to reduce the size of the FM optimization problem the full hydraulic model is simplified by the Simplifier, while retaining the nonlinear characteristics of the model. In the simplified model all reservoirs, elevated tanks and all control elements, such as pumps and valves, remain unchanged, but the number of pipes and nodes is

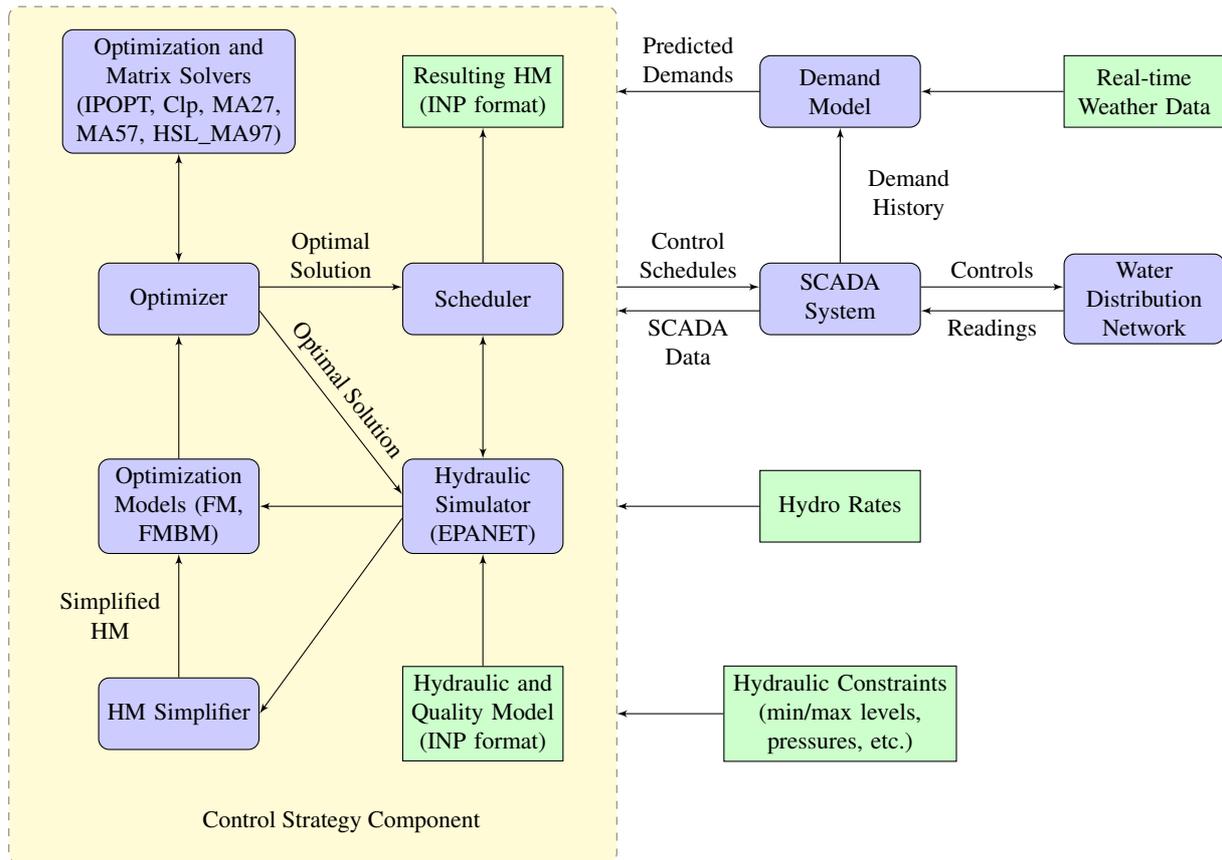


Figure 1: Structure of the TOO system and CSC component

significantly reduced. The final reservoir states for the 24-hour FM problem are taken from the solution of the full mass-balance model (FMBM). The FMBM model is a large-scale linear programming model, solved usually over weekly horizon, and emerges directly from the FM model for which all pressure dependent variables (heads, levels, head losses and head gains) and constraints were omitted or substituted by appropriate parameters. The EPANET hydraulic simulator is used to generate the starting point for IPOPT solver, to validate the optimal aggregated solution provided by FM optimization, and also by the Scheduler to adjust an optimized pump schedule by feedback mechanism from the simulated hydraulic model. The FM optimization model is solved by IPOPT optimizer to produce continuous optimal solution. When the CSC application is employed in real-time TOO environment, the optimal solution from the previous FM optimization can be used as an initial guess for the current FM step. The FM optimal solution is then fed back from IPOPT into the Optimizer for analysis, further processing and exporting of the results outside of the main software module. The CSC module is able to produce resulting EPANET INP file with optimized pump schedules and to send optimized controls (optimal pump schedules and the resulting reservoir level trajectories) to the SCADA system. Moreover, the CSC is also using short term demands predicted by the Demand Model; it works with different demand levels (scaled for different seasons). In general, the CSC runs as follows:

- 1) Collect external factors (weather, energy rates), system status and data. This includes, but is not limited to, current reservoir and elevated tanks levels, equipment out-of-service, equipment auto/manual modes, production costs.

- 2) Run demand model to predict demand.
- 3) Determine final states for reservoirs and elevated tanks by solving the FMBM problem over long-term horizon.
- 4) Determine optimized pump schedules and tank profiles by solving the FM problem over short-term horizon (with final states taken from the solved FMBM step), and application of the Scheduler.
- 5) Run hydraulic/quality model to check optimized pump schedules.
- 6) Analyze results. If results are acceptable (hydraulically feasible), then apply control strategy to SCADA system.

The majority of operational control approaches for water distribution networks (WDN) reported in the literature usually consider small-scale systems as case studies. Moreover, they are typically based on the simulation-optimization methodology using a hydraulic simulator, or they use a simplified mass-balance models as a key element of their optimization process. It is common for practically solved optimal control problems that an objective function is evaluated by solving a complex model (Arabas and Malinowski, 2001; Szynekiewicz and Błaszczyk, 2011), and quite often the only way to effectively deal with complexity of the optimization problem is to use the simulation-optimization approach (Niewiadomska-Szynekiewicz, 2004; Kamola, 2007).

To determine optimal controls the CSC Optimizer uses a full hydraulic model of the water transmission system. The FM optimization model reflects full hydraulic characteristics of the TWS and can be automatically adapted to structural changes in the network, such as isolation of part of the network or changing volume of reservoir by adding/removing a cell, as

well as to changes in operational constraints. Furthermore, both optimization models used by the CSC can be generated and solved in an automatic manner for different time horizons and configured by a large set of available options. A more detailed description of the Toronto Water System and the TOO project can be found in (Błaszczuk et al., 2012b,a, 2013).

FM OPTIMIZATION MODEL

The optimization of pump operations is a difficult task due to the significant complexity and non-linearity of WDNs, as well as due to the number of operational constraints and interactions between different network elements. For example, pumps usually do not operate in isolation – a change in the operating duty of one pump may affect the suction or discharge pressure of other pumps in the same hydraulic system.

Optimization methods described in this paper are model based. The hydraulic model is provided as the EPANET INP file and consists of:

- 1) boundary conditions (water sources, initial reservoir levels and demands),
- 2) a hydraulic nonlinear network made up of pipes, pumps, valves, and
- 3) reservoir dynamics.

The NLP algorithm that has been used to compute the continuous schedules, and also the discretized schedule, require the EPANET simulator of the hydraulic network. In the TOO system the EPANET Toolkit has been used to provide the initial feasible solution to the NLP solver by simulating the network, and also the Toolkit has been utilized by the TOO scheduler for discretization of the continuous solution.

The main objective is to minimize the pumping cost subject to the hydraulic network equations and operating constraints over a given time horizon with hourly discretization. The relationships between different components of cost and operating constraints in the whole hydraulic network are of a very complex nature, thus they can only be solved by use of an advanced nonlinear programming solver and optimal scheduler.

The goal of the optimal network scheduling is to calculate the least-cost operational schedules for pumps, valves, and water treatment plants for a given period, typically for 24 hours or one week. The optimization problem is expressed in discrete-time, i.e., in the FM model an hourly time-step is used. The FM optimization problem is given as:

- 1) minimize objective function consisting of pumping cost and water treatment cost,
- 2) subject to hydraulic network equations, and
- 3) operational constraints.

These three parts of the problem are briefly discussed in the following subsections. More detailed description and discussion for the FM model is given in (Błaszczuk et al., 2013).

Objective function

The objective function to be minimized is the sum of two costs associated with the system: pumping cost and water treatment cost. The pumping cost depends on the efficiency of the pumps used and the electricity power tariff over the pumping duration. The tariff is usually a function of time with cheap and more expensive energy periods. The water treatment cost for each water treatment plant (WTP) is proportional to the flow output from a WTP with the unit price.

In general, the pumping cost may be reduced by decreasing the water quantity pumped, decreasing the total system head,

increasing the overall efficiency of the pumping station by proper pump selection, or using reservoirs and elevated tanks to maintain highly efficient pump operations. In most instances, efficiency can be improved by using an optimization algorithm to select the most efficient combination of pumps to meet a given demand. Additional cost savings may be achieved by shifting pump operations to off-peak water-demand periods through proper filling and draining of reservoirs and elevated tanks. Off-peak pumping is particularly beneficial for systems operating under a variable-electric-rate schedule.

Decision variables

The decision variables in the resulting aggregated nonlinear optimization problem are the average aggregated flows and average head gains for all logical pumping stations at each hour of the control horizon. Also, the decision variables might be the settings for some throttled valves (minor losses or valve openings) and settings for pressure reducing valves (pressure set-points) in the hydraulic system. The indirect decision variables in the optimization problem are:

- flows and head losses for every pipe and valve
- heads at every junction and demand node
- heads, volumes and water levels for every reservoir and elevated tank

For all those variables there are simple bounds constraints. All variables are related mutually through the hydraulic model.

Hydraulic model

Each network element has a hydraulic equation. For pipes equations, the Hazen-Williams formula is used (Brdys and Ulanicki, 1994). In the optimal scheduling problem it is required that all calculated variables satisfy the hydraulic model equations. The network equations are usually non-linear and are embedded as inequality and equality constraints in the optimization problem. The hydraulic model used by the FM optimization model consists of the following network equations:

- flow continuity at connection nodes
- mass-balance, average head and volume curve for reservoirs and elevated tanks
- head-loss for pipes
- head-loss for TCV valves
- check valves
- PRV valves
- pumping stations

The hydraulic model for TOO system is very complex, non-linear and large scale. It includes many types of linear and non-linear constraints modeling behaviour of all network elements.

Operational constraints

The operational constraints have the form of simple inequalities and are applied to keep the system state within its allowed operating range. Thus, we must take into account time varying minimum and maximum reservoir and elevated tank levels and volumes. The reservoir and elevated tank volumes (state variables) should remain within the prescribed simple bounds in order to prevent emptying or overflowing, and to maintain sufficient storage for emergency purposes. The reservoir and elevated tank level constraints are not necessarily equal to the physical limits provided in the EPANET INP file. Similar constraints must be applied to the heads at critical nodes (SYPs) in order to maintain required pressures throughout the water network. The other variables, such as:

- flows for all links (pipes including CVs, TCV and PRV valves, and pumping stations)
- head-losses for pipes and valves, and head-gains for pumping stations
- heads at all nodes (connection junctions, demand nodes, suction and discharges of pumping stations)
- water levels for reservoirs and elevated tanks

are also constrained by lower and upper limits determined by the features of particular network elements.

Other important constraints are on the final water level (and final water volume) of reservoirs and elevated tanks, such that the final level is not smaller than the initial level. Without such constraints the least-cost optimization would result in emptying all reservoirs. In the case of TOO system such constraint is applied over a long-horizon (up to 7 days) when solving a mass-balance optimization problem.

Final states for reservoirs and elevated tanks

The objective function, representing the total operating cost to be minimized, is usually comprised of energy cost for pumping water and the cost for treating water, although other costs such as penalties for deviation from the final reservoir (and elevated tank) target levels are sometimes included. The final penalty charge is associated with the cost imposed on the state variables for deviation from the specified final reservoir levels.

HYDRAULIC SIMULATION BY EPANET

EPANET is a public domain software developed by the Water Supply and Water Resources Division of the U.S. Environmental Protection Agency's National Risk Management Research Laboratory (Rossman, 2000). EPANET provides an integrated environment for editing network input data, running hydraulic and water quality simulations, and viewing the results in a variety of formats. The hydraulic simulation performed by EPANET delivers information such as flows and head losses in links (pipes, pumps and valves), heads, pressures and demands at junctions, levels and volumes for water storage. This allows computing the pumping energy and cost. EPANET's computational engine is available also as a separate library (called the EPANET Toolkit) for incorporation into other applications. The network hydraulics solver employed by EPANET uses the Gradient Method, first proposed by Todini and Pilati (Todini and Pilati, 1988), which is a variant of Newton-Raphson method.

While EPANET is used as the computational engine for most water distribution system models, most models are developed and maintained in hydraulic modeling packages based on EPANET's computational engine. Some of the major hydraulic modeling packages are:

- InfoWater, developed by Innovyze (formerly MWH Soft, a subsidiary of MWH)
- MIKE URBAN, developed by DHI
- WaterCAD and WaterGEMS, developed by Bentley's Haestad Methods (Hydraulics & Hydrology) group

The EPANET Programmer's Toolkit is a software library of functions that allow developers to customize EPANET's computational engine for their own specific needs. The functions can be incorporated into an applications written in C/C++, Delphi Pascal, Visual Basic, or any other language that can call functions within the EPANET Toolkit library. There are over 50 functions that can be used to open a network description file, read and modify various network design and operating parameters, run multiple extended period simulations accessing

results as they are generated or saving them to file, and write selected results to file in a user specified format.

The EPANET Toolkit could be used for developing specialized applications, such as optimization or automated calibration models that require running network analyses as selected input parameters are iteratively modified. It was, therefore, decided that EPANET will be used as the hydraulic model component of the TOO system rather than InfoWater, since the EPANET Toolkit is capable of providing all of the hydraulic modelling functionality that InfoWater provides with the ability to automate the entire process.

InfoWater is used as the primary hydraulic modelling software package for the TWS. The package is capable of producing a text file of the EPANET text file input format (INP). The on-line TOO Optimizer uses EPANET for simulation purposes and the INP file is imported into the tool's structure, which in turn uses the information therein for simulation purposes. Creation and maintenance of this INP file can be handled using InfoWater and exporting the modified model from InfoWater to INP format, and then performing a comparable import at the on-line tool's end.

The modified and extended versions of the EPANET Toolkit and OOTEN library (C++ wrapper for EPANET Toolkit) have been built into the TOO Optimizer and are used to build the FMBM and FM optimization models, to generate the starting point for IPOPT optimizer and to check the aggregated optimal results provided by solving the FM problem. The EPANET Toolkit is also used during iterations of the TOO Scheduler to adjust the optimized pump schedule by feedback from hydraulic simulations of current iteration of the pumping schedule. Furthermore, it also provides the hydraulic and quality simulation results for the final optimized pump schedule, such as reservoir profiles for level volume and water quality, pressure profiles for PS discharges and SYP nodes, energy and power calculations for all pumping stations.

NONLINEAR PROGRAMMING BY IPOPT

The IPOPT solver is based on a primal-dual interior-point method (barrier method) used for nonlinear optimization relying on the solution of sequence barrier problems. The search direction is calculated in full or reduced space. The main advantages of IPOPT are: the possibility of solution of large-scale problems, the availability of different methods for calculation of search direction and for approximation of hessian of Lagrange function, various methods for the solution of reduced system of linear equations, the use in line-search minimization various merit functions and ensuring global convergence of the whole algorithm by using filter algorithm in line-search minimization.

The computationally most expensive part of the optimization algorithm implemented in the IPOPT solver (not including computations of the objective function, constraints and their derivatives) is the solution of the symmetric indefinite system of linear equations, which is most often of high order and has a sparse structure. For its factorization and solution, the IPOPT uses external sparse direct linear (matrix) solvers, such as MA27 (default option), MA57, HSL_MA77, HSL_MA86, HSL_MA97, WSMP, PARDISO and MUMPS.

For the interested reader, more information about IPOPT solver can be found in the doctoral dissertation (Wächter, 2002), in the article discussing in detail the primal-dual interior-point algorithm (Wächter and Biegler, 2006), and also at the web page <https://projects.coin-or.org/Ipop> (from which an open source C++ version of IPOPT is available).

In the TOO system, the network scheduling problem is solved by its implementation in C++ programming language and usage of the nonlinear programming solver IPOPT. The IPOPT solver was found to provide very good numerical performance, stability and robustness when solving the real-time NLP problems generated by the TOO system.

The IPOPT solver used in TOO was configured to use the HSL_MA97 matrix solver (Hogg and Scott, 2011, 2013), compiled with OpenMP support to allow for parallel matrix computations. The HSL_MA97 linear solver was found to offer very good performance and robustness for the solved FM problems. It is also bit-compatible, which means that running (in parallel) the same matrix factorization twice will result in the same answers. Such a feature is very important for testing and debugging purposes, i.e., to obtain the same result of optimization regardless of the used number of OpenMP threads.

CONTROL STRATEGY COMPONENT

The Control Strategy Component (CSC) within TOO has been implemented in C++ programming language by the use of a few auxiliary software components, including:

- extended and fixed version of the EPANET Programmers Toolkit 2.0.12
- OOTEN library (C++ wrapper for the EPANET Toolkit)
- COIN-OR IPOPT (Interior Point Optimizer) solver for NLP problems
- COIN-OR CLP solver for LP problems
- CppAD package for automatic differentiation of C++ algorithms
- BOOST C++ library
- GNU Scientific Library (GSL)
- matrix solvers MA57 and HSL_MA97 from Harwell Subroutine Library (HSL)
- Intel Math Kernel Library (MKL)
- Oracle C++ Call Interface (OCCI)

By solving of the FM problem, the CSC ensures that pre-set minimum and maximum (critical) storage levels are not violated and that optimal pumping strategies are achieved for different seasonal, weekday/weekend and peak-day demands, as well as when abnormal events occur (e.g., pumping station, filtration plant or reservoir cell out-of-service). It also considers the production cost of water which varies from plant to plant in developing the optimal solution. The CSC uses water demand forecasts (for each demand node) and the system hydraulic and water quality model (defined in EPANET INP format). The computed optimal control strategies (pumping schedules and reservoir profiles) enable optimization of water pumping and water quality in the Transmission System. The used FM optimization model is based on the full hydraulic model, thus the optimal FM results are always consistent with results of hydraulic simulation by EPANET. The optimal aggregated FM solution is always validated and analyzed by use of the hydraulic simulation. We have found in very many CSC testing runs (also for CSC working in on-line mode) that FM optimization model provides practically the same results as a hydraulic simulator.

NUMERICAL RESULTS

To solve the FM optimization problem by IPOPT solver both in real-time and in robust way, we had to implement a strategy to generate the starting point for IPOPT solver and also a scaling method of decision variables and NLP problem functions.

Selection of starting point

An important requirement for solving of NLP problems is the selection of a starting point for numerical iterations. Nonlinear programming is a local search method and the starting point should be as close as possible to the final solution. Since both IPOPT solver and EPANET Toolkit are integrated into the CSC software component with a common data structure, the EPANET simulator is used to provide an initial starting point for the network scheduling problem. This facilitates the solution of the initialization problem in a very efficient manner.

At first, the hydraulic simulation is performed using the historical pumping schedules from the initial INP file. The results of this hydraulic simulation (i.e., flows, heads, levels, volumes, head-losses and head-gains) are passed to the NLP solver as a starting point. The next step before starting the right optimization is an initialization of the short-horizon (usually one-day) FM problem by use of the data fetched from the TOO database such as initial reservoir and elevated tank levels, predicted nodal demands, predicted spot energy prices, and the final desired reservoir and elevated tank levels (taken from the optimal solution of the FMBM problem). Finally, all the initial values for decision variables are projected into a simple bound constraints, which are defined internally in the CSC, or are provided by the TOO (for levels, volumes and pressures at SYP points and at discharge nodes of pumping stations).

Scaling of optimization problem

IPOPT solver has some mechanisms to find internal scaling factors; the default one only tries to handle over-scaling, i.e., cases in which the derivatives at the initial point are very large. In general, it is advised to try to scale the optimization problem so that the nonzero elements in the objective and constraint function gradients are roughly on the order of 0.01 to 100 for the points of interest. However, finding a good scaling for an optimization is not an easy task. Obviously, it would be great if the optimization codes could do that on their own, but due to the nonlinear nature of the functions that is in general a tough call. A well-scaled problem makes the solution process easier for any nonlinear optimizer, not just IPOPT. In the case of TOO system we use scaling for flow variables in the constraints modeling CV and PRV valves, and also pumping stations. Furthermore, there is a scaling for volume variables in equations modeling reservoirs and elevated tanks.

Optimization results

The IPOPT solver calculated the optimal aggregated flows and head gains at each logical pumping station over time horizon of 24 hours (with hourly discretization). The FM problem was build from the simplified hydraulic model. Model size reduction, based on an elimination of dead links and short or low resistance pipes, merging of pipe series, parallel pipes and non-critical nodes, allows for much faster IPOPT convergence to a local optimal solution.

The resulting FM model is a large-scale nonlinear optimization problem. The basic, 24-hour period, version of FM model (with additional variables and constrains to deal with an infeasible initial and final states, and also with an infeasible pressure and volume limits) consists of over 100,000 decision variables and nearly 106,000 equality and inequality constraints. The 7-day FM problem has almost 642,000 variables and 680,000 constrains. To better show the obtained optimal FM solution, we present here aggregated results for the control horizon of 7 days. The aggregated original and optimized volume profiles (also simulated by EPANET to check their correctness) for

all reservoirs and elevated tanks are presented in figure 2. The figure shows also the total power usage for all pumping stations for manual and optimized aggregated pumping flows and head gains. As can be seen, the aggregated FM optimized volume profile (*Vol Optim*) and that simulated by EPANET (*Vol Optim-EPA*) are consistent. It was found that the FM model adequately replicates the hydraulic behaviour of the original hydraulic model provided in the EPANET INP format. The optimal aggregated volume profile has a direct correspondence with the electricity tariffs. The water storages are emptying during higher tariff periods and vice versa. Also, the optimized energy consumption (*Power Optim*) is higher during lower tariff periods. The obtained energy cost savings, compared to the manual pump controls, were between 5 and 15%.

The IPOPT optimization solver working with HSL_MA97 matrix solver was executed on an Intel i7 X980 CPU with a clock speed of 3.33 GHz. The solution times for the tested 24-hour FM problems were within 2-10 minutes (depending on the used boundary conditions) and required 200-1000 IPOPT iterations. Optimization for 7-day horizon with hourly time-step took around 1-2 hours with more than 1000 IPOPT iterations required. The starting point for IPOPT was generated from EPANET hydraulic simulation of TWS hydraulic model with historical pumping schedules. IPOPT was configured to use OpenMP version of the HSL_MA97 matrix solver and 6 OpenMP threads. The HSL_MA97 solver proved to be the fastest, most stable and reliable matrix solver for IPOPT when solving the FM problem. The IPOPT solver was found to provide very good performance, stability and robustness when solving real-time FM problems generated by the TOO system. It is seen from the Table 1 that optimization time required by IPOPT solver scales well with increasing FM problem size.

Table 1: Optimization results with IPOPT and OpenMP HSL_MA97 for the FM problems of different time horizons; N – number of hourly intervals of time horizon, n – number of decision variables, m – number of constraints, T – solution time (in seconds), I – number of IPOPT iterations, S – aggregated energy cost savings (in %).

N	n	m	T	I	S
1	4093	4164	4.44	223	5.36
2	7910	8213	1.62	51	5.92
3	11727	12262	2.61	63	6.27
6	23178	24409	5.01	62	7.43
12	46080	48703	29.55	166	15.61
18	68982	72997	61.45	209	15.36
24	91884	97291	104.62	225	14.80
36	137688	145879	336.14	430	16.56
48	183492	194467	1249.67	673	15.01
72	275100	291643	1748.46	573	15.65
96	366708	388819	2698.49	584	16.50
120	458316	485995	2675.75	710	15.89
144	549924	582707	2519.70	633	15.52
168	641532	679419	4538.85	1172	15.99

In fact, the authors have found so far only one truly comparable problem, in size and complexity, reported in the literature and concerned with operation of the Berlin Water Works (Berlin Wasserbetriebe), presented in (Burgschweiger et al., 2009b,a). Yet, TWS is about twice as big in size, i.e., in the number of components, than the Berlin system. Also, in the FM model used by TOO system there was a need to include PRV valves, and to use complicated energy tariffs. These components

were not present in the schedule optimization for Berlin WDS, while the inclusion of such elements increases considerably both the modeling effort and complexity of the optimization problem.

CONCLUSIONS

We described, in general, the concept of TOO system and a complex, large-scale non-linear FM optimization model, based on the system of hydraulic equations for all elements comprising the water distribution system. The FM optimization model is automatically obtained from a hydraulic model in EPANET format and from additional files describing operational constraints, electricity tariffs and pump station configurations. Then, the FM model is solved over a 24-hour control horizon to obtain an optimal aggregated flows and average pressure gains for all pumping stations. The obtained optimal volume profiles for all reservoirs and elevated tanks confirm a high accuracy of the obtained FM optimal solution when compared with results of hydraulic simulation under EPANET. Thus, the presented FM model can be used as a hydraulic engine embedded in the nonlinear optimization models used for operational control of water distribution systems.

REFERENCES

- Arabas, P. and Malinowski, K. (2001). Periodic coordination in hierarchical air defence systems. *International Journal of Applied Mathematics and Computer Science*, 11(2):493–513.
- Błaszczuk, J., Karbowski, A., Krawczyk, K., Malinowski, K., and Allidina, A. (2012a). Optimal pump scheduling for large scale water transmission system by linear programming. *Journal of Telecommunications and Information Technology (JTIT)*, 2012(3):91–96.
- Błaszczuk, J., Malinowski, K., and Allidina, A. (2012b). Aggregated pumping station operation planning problem (APSOP) for large scale water transmission system. In Jónasson, K., editor, *Applied Parallel and Scientific Computing, 10th International Conference, PARA 2010, Reykjavik, Iceland, June 6-9, 2010, Revised Selected Papers, Part I*, volume 7133 of *Lecture Notes in Computer Science*, pages 260–269, Berlin / Heidelberg. Springer-Verlag Inc.
- Błaszczuk, J., Malinowski, K., and Allidina, A. (2013). Optimal pump scheduling by non-linear programming for large scale water transmission system. In Callaos, N., Gill, T., and Sánchez, B., editors, *Proceedings of The International Conference on Complexity, Cybernetics, and Informing Science and Engineering (CCISE 2013), June 30 - July 6, 2013, Porto, Portugal*, pages 7–12, Winter Garden, Florida, U.S.A. International Institute of Informatics and Systemics (IIS), Member of the International Federation for System Research (IFSR).
- Brdys, M. A. and Ulanicki, B. (1994). *Operational Control of Water Systems: Structures, algorithms and applications*. Prentice Hall, New York.
- Burgschweiger, J., Gnädig, B., and Steinbach, M. C. (2009a). Nonlinear programming techniques for operative planning in large drinking water networks. *The Open Applied Mathematics Journal*, 3:14–28.
- Burgschweiger, J., Gnädig, B., and Steinbach, M. C. (2009b). Optimization models for operative planning in drinking water networks. *Optimization and Engineering*, 10(1):43–73.
- Hogg, J. and Scott, J. (2013). New parallel sparse direct solvers for multicore architectures. *Algorithms*, 6(4):702–725.
- Hogg, J. D. and Scott, J. A. (2011). HSL_MA97: a bit-compatible multifrontal code for sparse symmetric systems. Technical Report RAL-TR-2011-024, STFC Rutherford Appleton Laboratory, Harwell Oxford, Oxfordshire, UK.
- Kamola, M. (2007). Hybrid approach to design optimisation: Preserve accuracy, reduce dimensionality. *International Journal of Applied Mathematics and Computer Science*, 17(1):53–71.
- Methods, H. (2003). *Advanced Water Distribution Modeling and Management*. Haestad Press, Waterbury, CT USA, first edition.
- Niewiadomska-Szynkiewicz, E. (2004). Computer simulation of flood operation in multireservoir systems. *SIMULATION – Transactions of The Society for Modeling and Simulation International*, 80(2):101–116.
- Rossman, L. A. (2000). EPANET 2 users manual. Technical Report EPA/600/R-00/057, U.S. States Environmental Protection Agency, National Risk Management Research Laboratory, Office of Research and Development, Cincinnati, Ohio, USA.

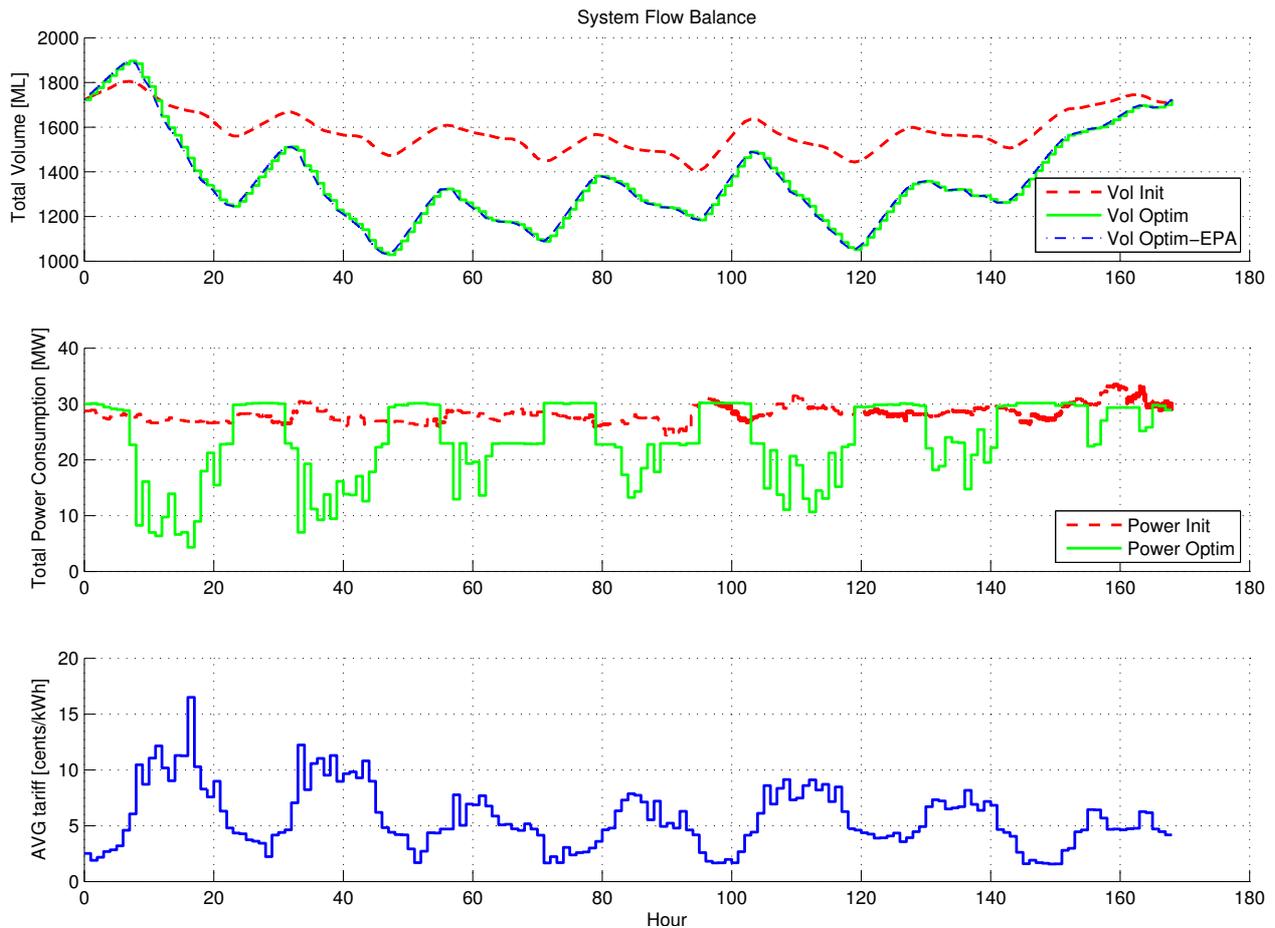


Figure 2: Aggregated volume profiles for all water storages of initial (*Vol Init*) and FM optimized solution – original (*Vol Optim*) and simulated by EPANET (*Vol Optim-EPA*); Total energy usage for all pumping stations of initial (*Power Init*) and optimized (*Power Optim*) pump schedules.

Szynkiewicz, W. and Błaszczuk, J. (2011). Optimization-based approach to path planning for closed chain robot systems. *International Journal of Applied Mathematics and Computer Science*, 21(4):659–670.

Todini, E. and Pilati, S. (1988). A gradient algorithm for the analysis of pipe networks. In Coulbeck, B. and Orr, C. H., editors, *Computer Applications in Water Supply: Vol. 1 – Systems Analysis and Simulation*, pages 1–20. Research Studies Press Ltd., Letchworth, Hertfordshire, England.

Wächter, A. (2002). *An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering*. Ph. D. dissertation, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

Wächter, A. and Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57.

AUTHOR BIOGRAPHIES



of large-scale water distribution systems.

JACEK BŁASZCZYK received his M.Sc. and Ph.D. degrees in Automatic Control from the Warsaw University of Technology, Poland, in 2000 and 2008, respectively. Currently, he is an Assistant Professor at the Research and Academic Computer Network (NASK). His research interest include large-scale nonlinear optimization, optimal control, parallel and distributed computations, numerical software for optimization and linear algebra, and recently, modeling, simulation and optimization



He was a visiting professor at the University of Minnesota; next he served as a consultant to the Decision Technologies Group of UMIST in Manchester (UK). Prof. K. Malinowski is also a member of the Polish Academy of Sciences.

KRZYSZTOF MALINOWSKI Prof. of Techn. Sciences, D.Sc., Ph.D., MEng., Professor of control and information engineering at Warsaw University of Technology, Head of the Control and Systems Division. Once holding the position of Director for Research of NASK, and next the position of NASK CEO. Author or co-author of four books and over 150 journal and conference papers. For many years he was involved in research on hierarchical control and management methods.



business in system automation, optimization and data management. The business is now part of IBI Group, and he is the Vice-President of IBI-MAAK Inc. He is responsible for technology development, business and strategic planning, and management. Current effort is focused on novel approaches to automation and system optimization with a focus on energy efficiency in real-time system control.

ALNOOR ALLIDINA received his B.Sc. (Hons) degree in Electrical and Electronic Engineering in 1977, and M.Sc. and Ph.D. degrees in 1978 and 1981 respectively, from the University of Manchester Institute of Science and Technology (UMIST), Manchester, UK. He held a tenured position with UMIST before taking on various industrial positions in the UK and Canada, focusing on the practical application of control theory. In 1991 he started a consulting and system integration

HYBRID CPU/GPU PLATFORM FOR HIGH PERFORMANCE COMPUTING

Michał Marks*, Ewa Niewiadomska-Szynkiewicz*,**

* Research and Academic Computer Network (NASK)

Wawozowa 18, 02-796 Warsaw, Poland

and

** Institute of Control and Computation Engineering

Warsaw University of Technology

Nowowiejska 15/19, 00-665 Warsaw, Poland

Email: mmarks@nask.pl, ewan@nask.pl

KEYWORDS

HPC, cluster, GPU computing, OpenCL, cloud computing, Openstack, cryptanalysis

ABSTRACT

High performance computing is required in a number of data-intensive domains. CPU and GPU clusters are one of the most progressive branches in a field of parallel computing and data processing nowadays. Cloud computing has recently emerged as one of the buzzwords in the ICT industry. It offers suitable abstractions to manage the complexity of large data processing and analysis in various domains. This paper addresses issues associated with distributed computational system and the application of mixed GPU&CPU technology to data intensive computation. We describe a hybrid cluster formed by devices from different vendors (Intel, AMD, NVIDIA). Two variants of software environment that hides the heterogeneity of our hardware platform and provides tools for solving complex scientific and engineering problems are presented and discussed. The first solution (HGCC) is a software platform for data processing in heterogeneous CPU/GPU clusters. The second solution (HGCVC) is an extension version of the previous one. The cloud technology is incorporated to the HGCC framework. The results of numerical experiments performed for parallel implementations of password recovery algorithms are presented to illustrate the performance of our systems.

INTRODUCTION

Computational-effective High Performance Computing (HPC) is required for efficient transformation of massive data into valuable information and meaningful knowledge. It is obvious that in order to support calculations for fast increasing number of data more sophisticated

software and hardware platforms to perform complex operations in a scalable way have to be developed.

In recent years parallel processing has provided a new impetus in systems engineering. The intense research, development and deployment of hardware, software and applications for parallel computers were carried out. During the 1980s and 1990s, software for parallel computers focused on providing powerful mechanisms for managing communication between processors, and environments for parallel machines and computer networks. High Performance Fortran (HPF), OpenMP, Parallel Virtual Machine (PVM) and Message Passing Interface (MPI) were designed to support communications for scalable applications (Karbowski and Niewiadomska-Szynkiewicz, 2009). The application paradigms were developed to perform calculations on shared and distributed memory machines.

In the last decade, clusters, grids and clouds have been identified as important new technologies for massive data processing and large scale computing. In today's computer systems these technologies are often used to solve complex scientific problems as well as to tackle projects in industry and commerce (Marks, 2012; Wang et al., 2011; Niewiadomska-Szynkiewicz and Marks, 2012; Szynkiewicz and Błaszczuk, 2011; Koodziej et al., 2013). The most common operating systems used for building clusters are UNIX and Linux. Clusters should provide scalability, transparency, reconfigurability, availability, reliability, and high performance. There are many software tools for supporting cluster computing, such as SLURM (Yoo et al., 2003), Torque/MOAB (Staples, 2006) or ASimJava (Sikora and Niewiadomska-Szynkiewicz, 2007). A novel approach to perform parallel computations is to use a hybrid cluster – the computational architecture with multicore CPUs working together with multicore GPUs (Kunzman and Kalé, 2011; Wen-Mei, 2011). Graphics Processing Units (GPUs) al-

low to perform massively parallel computations. Many operations are natively supported by GPU units, involving maximum instruction throughput and full use of computational resources. Using CUDA or OpenCL software platforms many applications can be easily implemented and significantly faster executed than on multiprocessor or multicore computational systems.

Initially the idea of Grid was to extend parallel computing paradigms from tightly coupled clusters to geographically distributed systems. However, in practice, Grid has been utilized more as a platform for the integration of loosely coupled applications (Berman et al., 2003; Kołodziej et al., 2014). Currently computational Grids enable the sharing and aggregation of a wide variety of geographically distributed computational resources, such as supercomputers, computer clusters, data sources, storage systems, scientific instruments, and present them as a unified, dependable resource for solving large-scale computations and data intensive computing applications. Grids have the potential to integrate as never before - theory, experiment, and computation - and to do so on a global scale.

Clouds are the natural evolution of traditional clusters and data centres (Wang et al., 2011, 2010). They are distinguished by pricing model where customers are charged based on their utilisation of computational resources, storage and transfer of data. The cloud computing emerges as a new computing paradigm that offers reliable, customized and QoS guaranteed dynamic computational environments. Clouds offer services to access hardware, software and data resources in a transparent way. The services are referred to as:

- Platform as a Service (PaaS),
- Software as a Service (SaaS),
- Infrastructure as a Service (IaaS).

Due to abilities to provide flexible computational infrastructure, configurable software services and pay-as-you-use cloud computing seems to be very promising in massive data processing and solving complex computing problems. These emerging services can reduce the cost of computation, application hosting and content storage and delivery by several orders of magnitude.

This paper addresses issues associated with distributed computational systems and the application of mixed GPU&CPU and cloud computing technologies to massively parallel computations. These technologies are very effective in solving complex calculation problems that can be divided up into large numbers of independent parts. The particular improvement should be obtained for problems that can be solved applying SPMD

(Single Program Multiple Data) paradigm. The data decryption/encryption and password recovery algorithms are easy adaptable to parallel environments. Therefore, GPU-based computation performed in cluster and cloud infrastructures can be especially exploited in the field of cryptanalysis and cryptography.

The paper is organized as follows. In Section 2 we describe an architecture of our heterogeneous cluster formed by two types of CPU and GPU units. The software framework for managing calculations on such type of cluster is presented in Section 3. Finally, in Section 4 we summarize results of numerical experiments. The considered case study is concerned with parallel implementation of selected cryptanalysis algorithms. The paper is concluded in Section 5.

HARDWARE INFRASTRUCTURE

The objective was to develop a computational system composed of heterogeneous devices and software framework for massive parallel computations. The expected functionalities were:

- effective computation of applications implementing MapReduce programming model,
- integration of computational devices with different architectures (from different vendors) into one transparent system,
- resistance and easy to use.

Our hybrid hardware platform is composed of 24 nodes that integrates two types of multi-core CPUs and GPUs:

- 12 nodes equipped with: Intel Xeon X5650, 2.66 GHz/3.06 GHz turbo, 6 cores with Hyper-Threading technology/ 12 threads, 6x256 L2, 12 MB L3 cache, and NVIDIA Tesla M2050, 448 CUDA cores, 384-bit memory bus.
- 12 nodes equipped with: AMD Opteron 6172, 2.1 GHz, 12 cores / 12 threads, 12x512 KB L2, 12 MB L3 cache, and AMD FirePro V7800, 1440 stream processors (equivalent of 288 CUDA cores), 256-bit memory bus.

The system architecture is presented in Fig. 1. All computational nodes equipped with CPU and GPU devices are supported by a dedicated master and storage nodes providing access to disk arrays and management capabilities. Communication in the cluster presented in Fig. 1 is organized using different interconnects: InfiniBand 4x QDR, 10 GbE and 1 GbE. Such excess network configuration allows us to separate communication connected

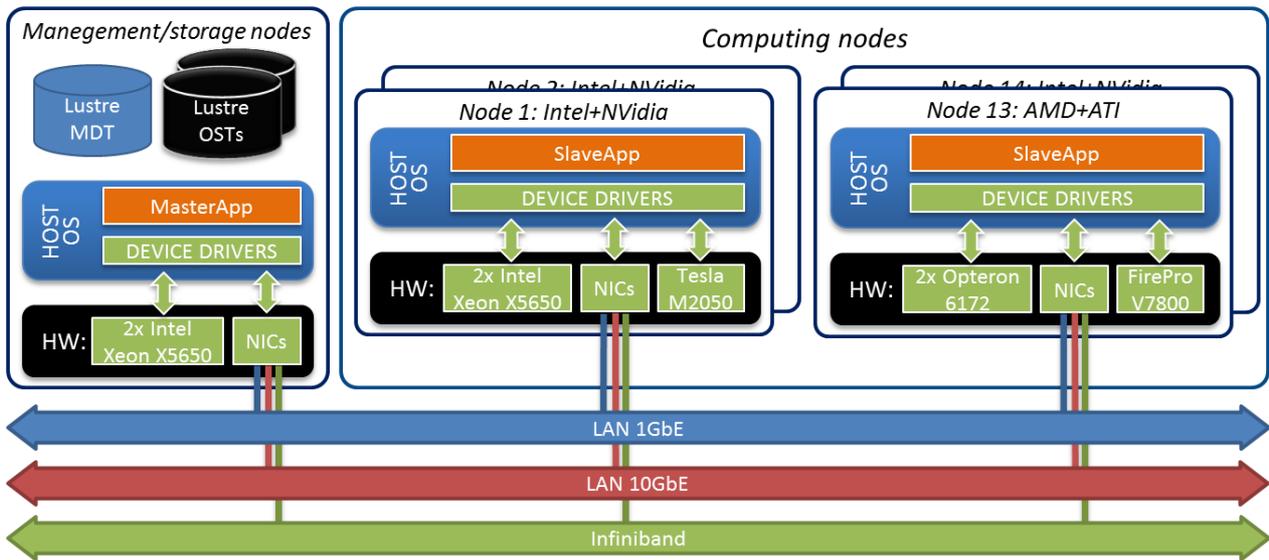


Figure 1: Hardware platform (Intel CPU + NVIDIA GPU and AMD CPU + AMD GPU nodes).

with IO operations from computational traffic. The current configuration assumes utilizing 10 GbE network for providing access to data storage. InfiniBand and the 1 GbE Ethernet are used for computational purposes.

SOFTWARE INFRASTRUCTURE

The novelty of our solution is not only the proposed hybrid architecture of our hardware platform but new software framework that can support the potential user in a task execution. The objective of this framework is to provide environments for parallel calculations that are performed on a hardware platform formed by heterogeneous CPU and GPU devices. The main functionality is to hide a heterogeneity of the computational nodes and minimize user's effort during the design, implementation and execution of the applications. We developed and implemented the software framework HGCC (Hybrid GPU/CPU Cluster) for management parallel calculations that can be performed on the hardware platform presented in Fig. 1. Moreover, this framework can be executed in virtual environment. HGCVC (Hybrid GPU/CPU Virtualized Cluster) is an extended version of HGCC utilizing virtualization and cloud computing. Application of Openstack orchestrator increases flexibility, functionality and robustness of HGCC software.

HGCC Framework

When designing the HGCC system we have assumed that from the potential user's perspective, the computational system should serve as one supercomputer. The concept

was to allow applications developed by users to transparently utilize many CPU and GPU devices, as if all the devices were on the local computer. A single system image model was implemented. Finally, in our framework all servers' resources such as CPU, GPU or memory are seen by the user as one unique machine. In order to take advantage of GPUs from different vendors, we decided to use OpenCL (Bainville, 2010). It is a low level programming toolkit for writing programs that execute across heterogeneous platforms consisting of CPUs, GPUs, DSPs (Digital Signal Processors), FPGAs (Field-Programmable Gate Arrays) and other processors.

In general, the goal of the HGCC framework is to divide the data into separate domains, allocate the calculation processes to cluster nodes, manage calculations and communication, and provide a store for all data objects. Hence, four groups of services are supplied: user interface services, calculation management services, communication services and data repository services.

HGCC is composed of several components. The most important are MasterApp (master node application) and SlaveApp (computational node application), Fig. 1. MasterApp is the main component that is responsible for the user-system communication and calculation management. SlaveApp is responsible for calculations that are performed by the assigned server.

It is assumed that each computational node in HGCC contains some number of resources. Two types of such resources are distinguished: CPUs and GPUs. The computational resource can be in one of three states:

- `waiting` – ready for loading a new task to execution,
- `working` – occupied, calculations are executed,
- `lost` – lost because of the node failure.

Our framework implements *master-slave* communication model. An XML-based communication protocol based on the TCP/IP protocol and BSD sockets is used to perform communication between master and slave nodes.

Each computational task should be defined in the `task descriptor` and implemented in an object oriented style. The XML Schema specification for building XML files with task description is provided in HGCC. The task descriptor contains: a type of the task, an algorithm, a destination platform and device. All these parameters are mandatory. Rest of this file is filled by parameters specific to a given task.

HGCC operates as follows. A computational task implemented by the user is sent to the `MasterApp` component. All parameters defined in the `task descriptor` are parsed inside `MasterApp`. The task is divided into smaller subtasks which are allocated to the slave nodes with free resources. Next the `SlaveApp` application is initialized. A plugin list is loaded from a plugin descriptor file, and a socket is opened for `MasterApp`'s connection. The plugin descriptor file contains information about all plugins currently available in the system. Whenever a slave node gets a new set of subtasks to execute it looks for available valid plugin, and loads it to the memory. Next, the control flow inside `SlaveApp` splits, and the newly spawned thread launches calculations stored in the loaded plugin.

HGCVC Framework

The HGCC framework is prepared to be executed in form of `MasterApp` and `SlaveApp` daemons working directly in the operating system. This solution is quite effective and comfortable when computational resources are dedicated to perform tasks supported by HGCC. However, in practice there are many situations with the need of assigning computational power to perform calculations on behalf of other projects. These situations often imply the necessity of preparing different environment and as a consequence it may lead to changes disrupting normal HGCC operations.

In order to overcome these problems we decided to prepare an extended solution utilizing the power of virtualization and cloud computing. The new environment is called HGCVC - Hybrid CPU/GPU Virtualized Cluster

and is based on KVM Hypervisor and Openstack orchestrator. The comparison of computational node architectures for HGCC and HGCVC is presented in Figure 2. In case of HGCC framework the `SlaveApp` is run by the host operating system (the only one available operating system) using hardware available through devices drivers. In case of HGCVC the `SlaveApp` is run by the guest operating system hosted on one of the Virtual Machines. The virtual machines are run by the hypervisor which provides the set of devices modules – a hardware abstraction. This property which is very useful in majority of cloud systems applications is a drawback in case of utilizing cloud computing for HPC. That's why in HGCVC solution the PCI pass-through property is exploited to provide a direct access to GPUs for guest operating systems (one guest OS in the same time). This solution allows us to minimize the overhead caused by virtualization and, in the same, allows us to get cloud benefits like scalability, reliability and utility of cluster infrastructure.

The whole process of orchestration and system management is organized using Openstack modules. In Fig. 2 only cloud agent (Openstack nova compute module) is presented – as this is the only module which needs to be run on computing nodes. The rest of Openstack functionality like network, storage, identity and images management is provided by the controller node which plays for Openstack a similar role like `MasterApp` for HGCC software.

CASE STUDY RESULTS: PASSWORD RECOVERY

However, the presented computational system composed of CPU and GPU devices can be used to any massively parallel and intensive-data computations we used it to develop efficient cryptanalysis and cryptography. The results of evaluations of selected encryption and decryption algorithms (DES, 3DES, AES) are described in (Niewiadomska-Szynkiewicz et al., 2012). In this paper we present and discuss the efficiency of parallel implementations of selected password recovery algorithms. The only reasonable technique for recovering a password from hash is to scan all potential password, compute their hash, and test the coincidence (Paar et al., 2010; Li et al., 2009). In general, cryptographic hash functions include integer and binary operations such as: addition modulo power of two, bit shift and rotation, bitwise xor, bitwise or, bit negation and words permutation. All those operations are natively supported by GPU processors. Three main approaches to password strength validation that are usually considered are: brute-force, rainbow-table and

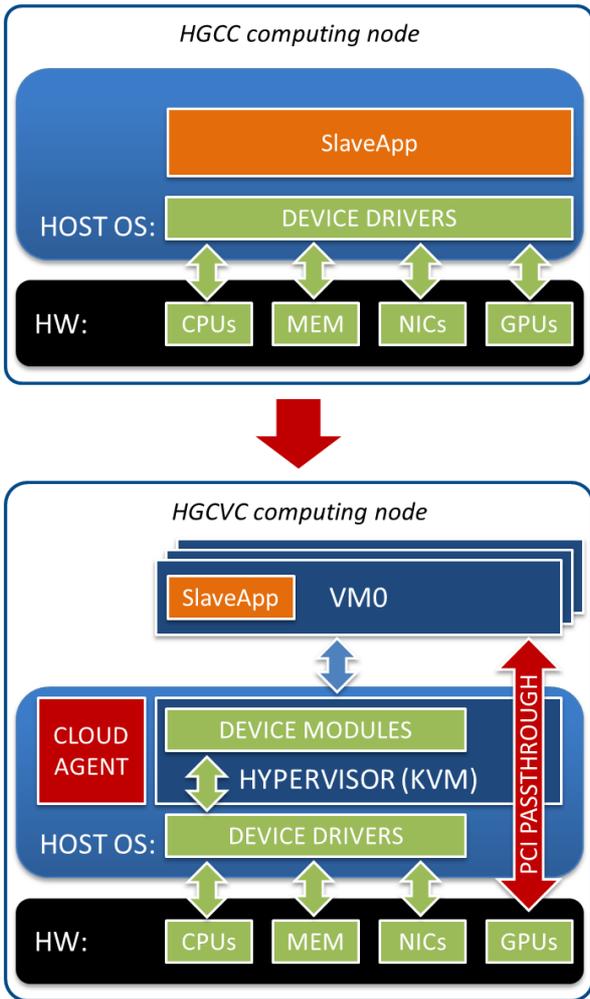


Figure 2: HGCC and HGVCV computational nodes.

dictionary test. The first two of these are very suitable for porting to GPU. Brute-force, thus not very sophisticated, is the most common approach used together with GPUs.

Multiple tests were performed for parallel implementations of password recovery from MD5, SHA-1 and SHA-2 hashes utilizing brute-force technique, and hybrid CPU&GPU computing. The aim of the first series of tests was to compare the performance achieved by Intel Xeon and AMD Opteron central processing units. Tables 1 - 3 present scalability of the CPU devices. Tables 1 and 2 collect the number of hashes generated per second using multi-threaded versions of respectively, MD5 and SHA-1 algorithms. The parameter th_num in all tables denotes the number of executed threads. The best results were obtained for Intel Xeon X5650 both in case of MD5 and SHA-1. In most tests two Intel Xeon processors with

th_num	1	2	4	8	16	24
AMD	20.1	39.8	79.6	157.7	315.2	465.4
Intel	37.8	74.5	147.9	284.2	384.7	407.5

Table 1: Number of generated MD5 hashes per second (in millions).

th_num	1	2	4	8	16	24
AMD	5.0	9.9	19.7	39.4	78.7	117.5
Intel	10.1	19.9	39.5	76.1	119.5	147.8

Table 2: Number of generated SHA-1 hashes per second (in millions).

6 cores each plus Hyper-Threading technology generated more hashes than two AMD Opteron processors with 12 cores each. Next, the scalability of both CPU technologies was compared. The results are presented in Table 3. It can be observed that both processors scale up very well with a certain predominance of the AMD device.

th_num	2	4	8	12	16	24
AMD (MD5)	2.0	4.0	7.9	11.8	15.7	23.2
Intel (MD5)	2.0	3.9	7.5	9.8	10.2	10.8
AMD (SHA-1)	2.0	4.0	7.9	11.8	15.8	23.5
Intel (SHA-1)	2.0	3.9	7.5	10.8	11.8	14.6

Table 3: Scalability of AMD Opteron 6172 and Intel Xeon X5650 processing units for MD5 and SHA-1 algorithms.

The goal of the second series of experiments was to compare the performance of CPU-based and GPU-based algorithms for password recovery. Three techniques for hash generation were considered: MD5, SHA-1 and four versions of SHA-2 (a-224 bit, b-256 bit, c-384 bit, d-512 bit). The number of hashes generated per second running MD5, SHA-1 and SHA-2 algorithms on Intel Opteron and NVIDIA Tesla processors are presented in Fig. 3. The results of the same tests performed on AMD Opteron and AMD FirePro processors are depicted in Fig. 4. As it can be seen in figures 3 and 4 the GPU-based implementations give better results than the CPU-based ones. In general, the best results were obtained for the OpenCL versions of the hash functions executed on AMD FirePro V7800.

Finally, we tested the scalability of the parallel implementations of MD5, SHA-1 and SHA-2 algorithms in the cluster. The aim was to present the efficiency of our hybrid computational system. The results, i.e., number of generated hashes per second are collected in Table 4. We present the results for subclusters composed of Intel Xeon, AMD Opteron, NVIDIA Tesla and AMD FirePro

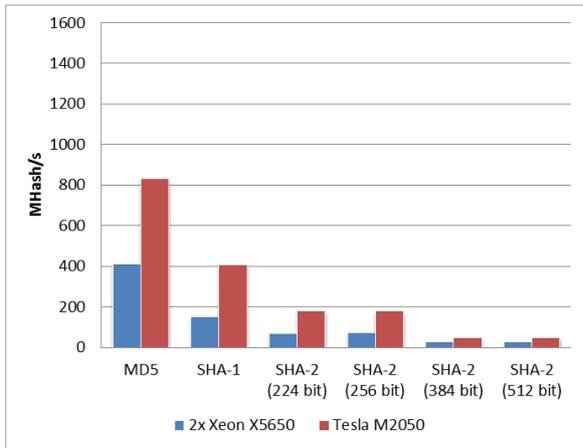


Figure 3: Number of generated hashes per second (in millions); Intel Xeon & NVIDIA Tesla.

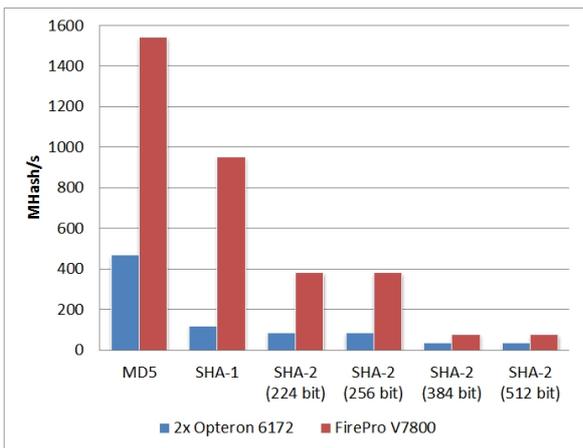


Figure 4: Number of generated hashes per second (in millions); AMD Opteron & AMD FirePro.

processing units. The last column contains the total results for the whole cluster. The results presented in Table

Algorithm	AMD Opteron	Intel Xeon	AMD FirePro	NVIDIA Tesla	Total
MD5	5518	4792	16599	10024	36933
SHA-1	1340	1750	11206	4952	19248
SHA-2a	945	791	4571	2094	8401
SHA-2b	940	793	4548	2094	8375
SHA-2c	417	297	859	532	2105
SHA-2d	410	298	870	529	2107

Table 4: Number of generated MD5, SHA-1 and SHA-2 hashes per second (in millions); the whole cluster.

4 show that our parallel implementations of password recovery algorithms scales up very well in the cluster composed of CPU and GPU devices.

SUMMARY AND CONCLUSION

In this paper we presented the short overview of the hybrid computational platform integrating CPU and GPU technologies. We focused on two variants of the software framework that integrate computational devices with different architectures into one transparent system. Our system has already proved to be very useful for massively parallel computing. The experimental results presented in the paper demonstrate its effectiveness and scalability. As a final observation we can say that cryptanalysis algorithms are natural candidates for massively parallel computing in computational platforms integrating CPU and GPU units. In our future research we plan to use HGCC and HGCVC to broad range of problems requiring large data processing.

REFERENCES

- Bainville, E. (2010). Opencl multiprecision tutorial @ONLINE.
- Berman, F., Hey, A., and Fox, G. (2003). *Grid Computing: Making the Global Infrastructure a Reality*. John Wiley & Sons, Inc., New York, NY, USA.
- Karbowski, A. and Niewiadomska-Szynkiewicz, E. (2009). *Parallel and distributed computing (in Polish)*. WUT Publishing House.
- Kołodziej, J., Khan, S., Wang, L., Kisiel-Dorohinicki, M., Madani, S., Niewiadomska-Szynkiewicz, E., Zomaya, A., and Xu, C.-Z. (2014). Security, energy, and performance-aware resource allocation mechanisms for computational grids. *Future Generation Computer Systems*, 31:77–92.
- Koodziej, J., Khan, S. U., and Talbi, E.-G. (2013). Scalable optimization in grid, cloud, and intelligent network computing foreword. *Concurrency and Computation: Practice and Experience*, 25(12):1719–1721.
- Kunzman, D. M. and Kalé, L. V. (2011). Programming heterogeneous clusters with accelerators using object-based programming. *Sci. Program.*, 19:47–62.
- Li, C., Wu, H., Chen, S., Li, X., and Guo, D. (2009). Efficient implementation for md5-rc4 encryption using gpu with cuda. In *Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference on*, pages 167–170. IEEE.

- Marks, M. (2012). Enhancing wsn localization robustness utilizing hpc environment. In *Proc. of the European Conference on Modelling and Simulation (ECMS 2012)*, pages 167–170. European Council for Modelling and Simulation.
- Niewiadomska-Szynkiewicz, E. and Marks, M. (2012). Software environment for parallel optimization of complex systems. In *Applied Parallel Scientific Computing K. Jonasson*, volume LNCS 7133, pages 86–96. Springer-Verlag.
- Niewiadomska-Szynkiewicz, E., Marks, M., Jantura, J., and Podbielski, M. (2012). A hybrid cpu/gpu cluster for encryption and decryption of large amounts of data. *Journal of Telecommunications and Information Technology*, 3(2):55–64.
- Paar, C., Pelzl, J., and Preneel, B. (2010). *Understanding Cryptography: A Textbook for Students and Practitioners*. Springer.
- Sikora, A. and Niewiadomska-Szynkiewicz, E. (2007). A federated approach to parallel and distributed simulation of complex systems. *International Journal of Applied Mathematics and Computer Science ACMS*, 17(1):99–106.
- Staples, G. (2006). Torque resource manager. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC '06*, New York, NY, USA. ACM.
- Szynkiewicz, W. and Błaszczak, J. (2011). Optimization-based approach to path planning for closed chain robot systems. *International Journal of Applied Mathematics and Computer Science ACMS*, 21(4):659–670.
- Wang, L., Ranjan, R., and Chen, J. and Beanatallah, B. e., editors (2011). *Cloud computing methodology, system, and applications*. CRS Press, Taylor & Francis.
- Wang, L., von Laszewski, G., Younge, A., He, X., Kunze, M., Tao, M., and Fu, C. (2010). Cloud computing a perspective study. *New Generation Comput.*, 28(2):137–146.
- Wen-Mei, W., H. (2011). *GPU Computing Gems Emerald Edition*. Morgan Kaufman.
- Yoo, A., Jette, M., and Grondona, M. (2003). Slurm: Simple linux utility for resource management. In Feitelson, D., Rudolph, L., and Schwiegelshohn, U., editors, *Job Scheduling Strategies for Parallel Processing*, volume 2862 of *Lecture Notes in Computer Science*, pages 44–60. Springer Berlin Heidelberg.

AUTHOR BIOGRAPHIES

MICHAŁ MARKS received his M.Sc. in computer science from the Warsaw University of Technology, Poland, in 2007. Currently he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2007 with

Research and Academic Computer Network (NASK). His research area focuses on wireless sensor networks, global optimization, distributed computation in CPU and GPU clusters, decision support and machine learning. His e-mail is mmarks@nask.pl

EWA NIEWIADOMSKA-SZYNKIEWICZ DSc (2005), PhD (1995), professor of control and information engineering at the Warsaw University of Technology, head of the Complex Systems Group. She is also the Director for Research of Research and Academic Computer Network (NASK). The author and co-author of 3 books and over 140 papers. Her research interests focus on complex systems modeling, optimization, control and simulation, parallel computation and computer networks. Her email is ewan@nask.pl.

USING ARTIFICIAL NEURAL NETWORK FOR MONITORING AND SUPPORTING THE GRID SCHEDULER PERFORMANCE

Daniel Grzonka and Joanna Kołodziej
Institute of Computer Science
Faculty of Physics, Mathematics and Computer Science
Cracow University of Technology
Warszawska st 24, 31-155 Cracow, Poland
E-mail: grzonka.daniel@gmail.com,
jokolodziej@pk.edu.pl

Jie Tao
Steinbuch Center for Computing
Karlsruhe Institute of Technology
Postfach 6980, D-76128 Karlsruhe, Germany
E-mail: jie.tao@kit.edu

KEYWORDS

Computational Grid, Scheduling, Artificial Neural Network, Data Classification, Security, Genetic Algorithm.

ABSTRACT

Task scheduling and resource allocations are the key issues for computational grids. Distributed resources usually work at different autonomous domains with their own access and security policies that impact successful job executions across the domain boundaries. In this paper, we propose an Artificial Neural Network (ANN) approach for supporting the security awareness of evolutionary driven grid schedulers. Making a prior analysis of the trust levels of resources and security demand parameters of tasks, the neural network monitors the scheduling and task execution processes. In the result produce the tasks-machines mapping “suggestions”, which can be then utilized by the scheduler to reduce the makespan or increase the system throughput. In this paper, we report the development of risk-resilient genetic-based schedulers and their integration with an ANN module of the HyperSim-G Grid Simulator to evaluate the proposed model under the heterogeneity and large-scale system dynamics. The simulation results showed a significant impact of the ANN support on enhancing the effectiveness of the genetic-based meta-heuristics in reducing the cost of security awareness in grid scheduling.

INTRODUCTION

Grid computing is one of the most popular combinations of traditional distributed computing and utility computing. This combination has become very effective for solving large-scale complex problems from various fields.

While the maximization of the resource utilization and profits of the resource owners are the key objectives of the grid scheduling, they may conflict with grid users’ security requirements and system reliability. A major hurdle in effective job outsourcing in grid is caused by network security threats. Therefore, it is desirable to have prior knowledge about the security demands from grid jobs and the trust level assured by a resource

provider at the grid cluster. An effective grid scheduler must be then security-driven and resilient in response to all scheduling and risky conditions.

The specific problem discussed in this paper is the improvement of the effectiveness of grid users’ and schedulers’ decisions on the low-cost resource allocations with security condition recognized as the most important factor. The security awareness of the grid schedulers in our approach is supported by an Artificial Neural Network (ANN) module which is monitoring the scheduling and task execution processes. The neural network learns patterns in input data (initial tasks and machines characteristics) and produces task-machine mapping recommendations based on the stored system information, such as resource failure rates and system input parameters. Thereafter, based on the ANN “suggestions”, sub-optimal schedules are generated and used in the initialization procedures of genetic-based schedulers, which may optimize the values of the main schedulers’ performance metrics (in our case makespan and flowtime). Despite the generation of the sub-optimal solution to the specified scheduling problems, the ANN module is not considered in this work in fact as another (separate) scheduler. It works in a “background” of the main process and monitors the scheduling results. However, the ANN’s outputs may be accepted as the optimal results by the employed scheduler.

Our contributions are summarized as the following:

- The development of an ANN-based model for supporting decision-making activities of grid actors in “secure” scheduling.
- The development risk-resilient genetic-based schedulers.
- The integration of ANN module and risk-resilient schedulers with grid simulator for their experimental evaluation.

This work considers the independent task model, where tasks submitted by the grid users are processed in a batch mode (Xhafa et al. 2007a). To incorporate security, we modified the Expected Time to Compute (ETC) model by integrating the security requirements as additional scheduling criteria. The scheduling problem

is defined as a bi-objective global minimization task with makespan and flowtime as the main scheduler's performance measures.

We evaluate the proposed ANN model and risk-resilient schedulers under the varying heterogeneity and large-scale system dynamics by using the *HyperSim-G Grid Simulator* (Xhafa et al. 2007b). We extended the previous version of *HyperSim-G* by an ANN module, in which the *Minimal Completion Time (MCT)* algorithm is used for the generation of sub-optimal schedules.

MODELS

In this section, we first define the main scheduling attributes that are necessary for the specification of a particular scheduling problem in CGs. Thereafter, we discuss a general model of the system for a security-aware management of tasks and resources.

Scheduling Problems' Model

There are four main scheduling attributes that must be specified to define a particular tasks-machines mapping problem, namely: (a) the environment, (b) grid architecture, (c) task processing policy, and (d) tasks' interrelations.

We consider in this work an *Independent Batch Scheduling in Hierarchical CG* problem, where it is assumed that the tasks are grouped into batches and can be executed independently in a hierarchically structured dynamic grid environment.

System Model

The architecture of a CG is usually modelled as a multi-level large-scale hierarchical structure, which is a compromise between centralized and decentralized resource and task managements. This hierarchy consists of three levels: (i) meta-scheduler, (ii) local task dispatchers and (iii) clusters.

A central meta-scheduler is the main module of the system working at the highest level. The meta-scheduler interacts with local task dispatchers (brokers) and CG users to generate optimal schedules.

The brokers collect information about the "computing capacities" of the resources supplied by the resource owners within the clusters, moderate the resources, and send all of the data to the meta-scheduler. The meta-scheduler must conceive an optimal plan of the resource allocations according to the various user requirements. Thereafter, replicas of the defined schedule are sent back to the brokers.

The role of the meta-scheduler is different when security is considered as an additional criterion within the scheduling process. The meta-scheduler must analyze the security requirements for the execution of

tasks and requests of the CG users for trustful resources available within the system. The system brokers analyze "reputation" indexes of the machines received from the resource managers and send proposals to the scheduler.

STATEMENT OF THE PROBLEM

We start the formalization of the security aware independent batch scheduling problem by introducing the following notation for tasks and computational resources, which will be used throughout the paper:

- n – the number of tasks in a batch;
- m – the number of machines available within the system for an execution of a given batch of tasks;
- $N = \{t_1, \dots, t_n\}$ – set of tasks within a batch;
- $M = \{x_1, \dots, x_m\}$ – set of available machines for the task batch;
- $N_i = \{1, \dots, n\}$ – labels of tasks;
- $M_i = \{1, \dots, m\}$ – labels of machines.

Characteristics of Tasks and Machines

Formally, each task j can be represented by a pair of parameters $j = (wl_j, sd_j)$, where:

- wl_j is a computational load of j expressed in Millions of Instructions Per Second (MIPS), we denote by $WL = [wl_1, \dots, wl_n]$ a workload vector for all tasks in the batch;
- sd_j is a security demand parameter, which is a component of a security demand vector $SD = [sd_1, \dots, sd_n]$. The major attributes affecting the security demand are data integration, task sensitivity, peer authentication, access control and task execute environment.

The workload of each of the submitted task can be estimated based on the specifications provided by the users, historical data, or it can be obtained from system predictions (Hotovy 1996).

Each machine i ($i \in M_i$) in the system is represented as a triplet $i = (cc_i, ready_i, tl_i)$, where:

- cc_i – is a computing capacity of i expressed in Millions of Instructions Per Second (MIPS), we denote by $CC = [cc_1, \dots, cc_m]$ a *computing capacity vector*;
- $ready_i$ – is a ready time of i , which expresses the time needed for the reloading of the machine i after finishing the last assigned task, a *ready times vector* for all machines is denoted by $ready_times = [ready_1, \dots, ready_m]$; and
- tl_i – is a trust level parameter, which specifies how much a grid user can trust a given resource manager; the manager maintains machine i status and monitors the execution of the tasks assigned to this machine; tl_i parameter

is the component of a *trust level vector* $TL = [tl_1, \dots, tl_m]$. The biggest impact on the parameter are prior task execution success rate, cumulative grid cluster utilization, capabilities of firewall, intrusion detection and intrusion response.

The detailed architectural structure of CG resources is not discussed in our approach. Therefore, it must be understood that the term “machine” in our system can be a single or multiprocessor computing unit or even refer to a local small-area network. However, we assume that: (a) a task can only be executed at one CG node in each batch; (b) no preemptive process is allowed within tasks or resources; (c) when a machine fails, tasks will be reallocated to other machine(s) in the next batch; (d) when a machine processes tasks, there is no priority distinctions between the tasks assigned in the previous batches and those assigned in the current batch; and (e) a machine must remain idle when tasks have been assigned to it.

We base our approach on the fuzzy-logic trust model (developed by Song et al. 2005). In this model the task security demand is supplied by the user programs as a single parameter. The demand may appear as request for authentication, data encryption, access control, etc.

The values of the sd_j and tl_i parameters are real fractions within the range $[0, 1]$ with 0 representing the lowest and 1 the highest security requirements for a task execution and the most risky and fully trusted machine, respectively. A task can be successfully completed at a resource when a *security assurance condition* is satisfied. That is to say that $sd_j \leq tl_i$ for a given (j, i) task-machine pair.

Let us denote P_f to be a *Machine Failure Probability* matrix, the elements of which are interpreted as the probabilities of failures of the machines during the tasks executions due the high security restrictions. These probabilities denoted by $P_f[j][i]$ are calculated by using the negative exponential distribution function as follows:

$$P_f[j][i] = \begin{cases} 0, & sd_j \leq tl_i \\ 1 - e^{-\alpha(sd_j - tl_i)}, & sd_j > tl_i \end{cases}$$

where α is interpreted as a failure coefficient and is a global parameter of the model.

Schedule Encoding

We use in our approach two different encoding methods of schedules, which can be defined as follows.

Definition: Let us denote by \mathcal{S} the set of all permutations with repetition of the length n over the set of machine labels M_i . An element $s \in \mathcal{S}$ is termed a *schedule* and it is encoded by the following vector:

$$s = [i_1, \dots, i_n]^T, \quad (1)$$

where $i_j \in M_i$ denotes the number of machine on which the task labelled by j is executed. The cardinality of \mathcal{S} is m^n .

This encoding method is called a *direct representation* of the schedule.

The direct representation of the schedules can be easily transformed into a *permutation-based representation*, in which, for each machine, a sequence of tasks assigned to that machine is specified. The tasks in the sequence are sorted (in increasing order) with respect to their completion times. Thereafter, all of the task sequences are concatenated into a vector u , which is in fact a permutation without repetition of tasks to machines. Formally, this kind of schedule encoding method can be defined in the following way:

Definition: Let us denote by $\mathcal{S}_{(i)}$ the set of all permutations without repetitions of the length n over the set of task labels N_i . A permutation $u \in \mathcal{S}_{(i)}$ is called a *permutation-based representation* of a schedule in CG and can be defined by the following vector:

$$u = [u_1, \dots, u_n]^T,$$

where $u_i \in N_i$, $i = 1, \dots, n$. The cardinality of $\mathcal{S}_{(i)}$ is $n!$.

In this representation some additional information about the numbers of tasks assigned to each machine is required. Therefore, we defined a vector $v = [v_1, \dots, v_m]^T$ of the size m , where v_i denotes the number of tasks assigned to the machine i .

Optimization Criteria and Objective Function

For estimating the execution times of tasks on machines we based our idea on the *Expected Time to Compute (ETC)* matrix model (Ali et al. 2000). The main structure in this model is the *ETC* matrix with estimated time needed for the completion of the task j on machine i .

In the simplest case, the entries of the $ETC[j][i]$ parameters can be computed as the ratio of the coordinates of WL and CC vectors. That is to say:

$$ETC[j][i] = \frac{wl_j}{cc_i}$$

The problem of scheduling tasks in can be measured under several criteria. In this work we consider the scheduling in CGs as a bi-objective global optimization problem with the hierarchical procedure of the minimization of *makespan* and *flowtime* objectives with *makespan* as a privileged criterion.

Using the *ETC* matrix model we can express the *makespan* and *flowtime* in terms of the completion times of the machines. The time of finishing the last task can be interpreted as the maximal completion time of the machines. Let us denote by *completion* a vector

of the size m , which indicates the time that machine i finalizes the processing of the previously assigned and planned tasks. That is to say:

$$completion[i] = ready_i + \sum_{\substack{j \in N_l: \\ s[j]=i}} ETC[j][i] \quad (2)$$

where $s \in Schedules$ denotes the schedule vector defined by Eq. (1).

The makespan can be now expressed as:

$$makespan = \max_{i \in M_l} completion[i] \quad (3)$$

We calculate the flowtime of the sequence of tasks on a given machine i by using the following formula:

$$flowtime[i] = ready_i + \sum_{\substack{j \in Sort[i]: \\ s[j]=i}} ETC[j][i] \quad (4)$$

where $Sort[i]$ denotes the set of tasks assigned to the machine i sorted in ascending order according to their ETC values.

Both makespan and flowtime are expressed in arbitrary time units. In fact, the numerical values are in incomparable ranges: flowtime has a higher magnitude order over makespan and its values increase as more jobs and machines are considered. Therefore, in this approach we use $mean_flowtime = flowtime/m$ for the evaluation of the flowtime criterion.

Based on the equations (2), (3) and (4) the two-steps optimization procedure can be defined as follows:

- **step 1:** minimize the maximal completion time of machines, i.e.

$$makespan = \max_{i \in M_l} completion[i] \rightarrow makespan_{min}$$

- **step 2:** minimize the flowtime without increasing the optimal makespan value, i.e.

$$current_makespan \leq makespan_{min}$$

Secure Mode Scheduling Scenario

In this paper we consider secure mode approach – when scheduler analyzes the *Machine Failure Probability* matrix in order to minimize the failure probabilities for task-machine pairs. In this scenario all of the security and resource reliability conditions are verified for the task-machine pairs. The main aim of the meta-scheduler is to design an optimal schedule for which, beyond the makespan and flowtime, the probabilities of failures of the machines during the tasks execution will be minimal. We assume that additional “cost” of the verification of security assurance condition for a given task-machine pair: (a) may delay the predicted execution time of the task on the machine and (b) is proportional to the probability of failure of the machine during the task execution. We define this “cost” as a

product $P_f[j][i] \cdot ETC[j][i]$ and the completion time of the machine i can be calculated as follows:

$$completion[i] = ready_time[i] + \sum_{\{j \in Tasks_i\}} (1 + P_f[j][i]) ETC[j][i]$$

where $Tasks_i$ denotes a set of tasks assigned to the machine i in a given batch.

ARTIFICIAL NEURAL NETWORK MODULE

For an ANN implementation we first provide a prior classification of tasks and machines available in the system based on the values of the WL , CC , TL and SD vectors. Machines are classified by the processing power (R_r classes: slowest, slower, ..., medium, ..., fastest) and the trust level (R_s number of classes: secure, less-secure, ..., medium, ..., fully-risky), where R_r and R_s are the parameters of the simulator. After the initial classification, the resources are divided into $R = R_r \cdot R_s$ classes (slowest-secure, ..., medium-secure, ..., fastest-fully-risky). We perform similar classification for the submitted tasks with workload and security demand instead of processing power and security criteria. We divide tasks into $T = T_w \cdot T_{sd}$ classes, where T_w is number of workload classes and T_{sd} is number of security demand classes. R machine classes and T task classes give us $R + T$ possible inputs for neural network.

To classify the output we need the results of monitoring the machine failures and successful execution of tasks on machines during the execution of the schedulers. We consider just the tasks which are successfully executed on the machines and then, for each machine class k we select the unique *major class* of tasks $\overline{t(k)}$, which contains the greatest number of successfully executed tasks.

The network is trained by the *back propagation algorithm* (Haykin 1999) and the generated outputs are used to compose the suboptimal schedules, which are moved to the initial population of the GA-based scheduler. We implemented the *Minimum Completion Time (MCT)* algorithm for generating those suboptimal schedules with a minimal completion time. The general framework of MCT is presented in Alg. 1.

Algorithm 1 MCT algorithm template.

- 1: Calculate the *ready_times* of the machines;
 - 2: **for all** Tasks in a given batch **do**
 - 3: Calculate the completion times of the machines for the tasks;
 - 4: Find the machine that gives minimum completion time, m_{best} ;
 - 5: Assign task to m_{best} machine;
 - 6: Update the machine completion time;
 - 7: **end for**
 - 8: **return** The resulting schedule;
-

SECURITY AWARE GENETIC-BASED BATCH SCHEDULERS

Heuristic methods are well known from their robustness and have been applied successfully to solve scheduling problems and general combinatorial optimization problems in dynamic environments (Aguirre and Tanaka 2007, Abraham et al. 2000, Kołodziej and Xhafa 2011). Therefore they can be considered as good candidates to be the effective CG schedulers that tackle the various scheduling attributes and additional security aspects.

In this work, we consider four genetic-based risk resilient grid schedulers presented in Table 1 that work in *secure* scheduling scenario.

Table 1: Four GA-based grid schedulers being evaluated

Scheduler	Replacement method	Scheduling scenario
GA-SS-S	Steady State	Secure
GA-SS-ANN	Steady State	Secure with ANN
GA-ST-S	Struggle	Secure
GA-ST-ANN	Struggle	Secure with ANN

Algorithm 2 A template of the genetic engine for four genetic-based grid schedulers.

```

1: Generate the initial population  $P^0$  of size  $\mu$ ;  $t = 0$ 
2: Evaluate  $P^0$ ;
3: while not termination-condition do
4:   Select the parental pool  $T^t$  of size  $\lambda$ ;  $T^t := Select(P^t)$ ;
5:   Perform crossover procedure on pairs of individuals in  $T^t$  with probability  $p_c$ ;  $P^t_c := Cross(T^t)$ ;
6:   Perform mutation procedure on individuals in  $P^t_c$  with probability  $p_m$ ;  $P^t_m := Mutate(P^t_c)$ ;
7:   Evaluate  $P^t_m$ ;
8:   Create a new population  $P^{t+1}$  of size  $\mu$  from individuals in  $P^t$  and  $P^t_m$ ;  $P^{t+1} := Replace(P^t; P^t_m)$ 
9:    $t := t + 1$ ;
10: end while
11: return Best found individual as solution;

```

We apply the *direct representation* of the schedules in the base populations P^t and P^{t+1} , and *permutation representation* in P^t_c and P^t_m populations. In the algorithms the initial population is generated by using the *MTC + LJFR-SJFR* method, in which all but two individuals are generated randomly. Those two individuals are created by using the *Longest Job to Fastest Resource – Shortest Job to Fastest Resource (LJFR-SJFR)* and *Minimum Completion Time (MCT)* heuristics (Xhafa et al. 2007b). In *GA-SS-ANN* and *GA-ST-ANN* algorithms the initial populations contain additionally the schedule generated by ANN module (by using the MCT heuristics).

The aforementioned methodologies differ in the implementation of the replacement mechanism in the

main genetic framework. We used *Steady State* replacement in *GA-SS-xxx* algorithms and *Struggle* procedure in *GA-ST-xxx*.

In the *Steady State* method, a set of the best offspring replaces the worst solutions in the old base population. The main drawback of this method is that it can lead to premature convergence on some solution and impacts on the stagnation of the population. However, the aforementioned may be very fast in the fitness reduction.

A *Struggle* replacement mechanism can be an effective tool for avoiding too fast scheduler's convergence to the local optima. In this method, new generation of individuals is created by replacing a part of the population by the most similar individuals – if this replacement minimizes the fitness value.

The *Struggle* strategy has shown to be very effective in solving several large-scale multiobjective problems (see e.g., Bartschi Wall 1996, Grueninger 1997). However, the computational cost can be very high, because of the need of calculation of distances among all offspring in resulting population and the individuals in the base population for the current GA loop. To reduce the execution time of the struggle procedure we use a *hash technique*, in which the hash table with the *task-resource allocation* key is created. The value of this key, denoted by K , is calculated as the sum of the absolute values of the subtraction of each position and its precedent in the direct representation of the schedule vector (reading the schedule vector in a circular way). The hash function f_{hash} is defined as follows:

$$f_{hash}(K) = \begin{cases} 0, & K < K_{min} \\ \left\lfloor N \cdot \left(\frac{K - K_{min}}{K_{max} - K_{min}} \right) \right\rfloor, & K_{min} \leq K < K_{max} \\ N - 1, & K \geq K_{max} \end{cases}$$

where K_{min} and K_{max} correspond respectively to the smallest and the largest value of K in the population, and N is the population size.

EXPERIMENTAL EVALUATION

In this section we present the results of the experimental evaluation of four genetic-based schedulers defined in Table 1 in dynamic grid environment by using the grid simulator. The main aim of our simple analysis is to compare the effectiveness of the proposed metaheuristics in scheduling scenario and to verify the impact of the activation of ANN module on the failures of the machines as well as on the makespan and flowtime optimization.

Security Aware HyperSim-G Grid Simulator

To simulate the secure scheduling we define a *Secure HyperSim-G* simulator by extending the HyperSim-G framework. HyperSim-G simulator is based on a discrete event model. The sequence of events and the

changes in the state of the system capture the realistic grid dynamics. The simulator provides the full simulation trace by indicating a parameter for the trace generation. This functionality is useful for an easy implementation of the Neural Network module.

Based on the ETC Matrix model the Secure HyperSim-G simulator generates an instance of the scheduling problem by using the following input data: (i) the trust level vector TL , (ii) the security demand vector SD , (iii) the workload vector of tasks WL , (iv) the computing capacity vector of machines CC , (v) the vector of prior loads of machines $ready_times$, and (vi) the ETC matrix. The Neural Network module is designed for supporting the resolution methods used in the *Scheduler* class of the simulator. The output of the network is used to define a suboptimal schedule, which is copied to the initial population of a GA-based scheduler.

Experimental Settings and Performance Metrics

The Secure HyperSim-G simulator is highly parameterized to reflect the various realistic grid scenarios. The values of key input parameters for the simulator are presented in Table 2.

Table 2: Values of key parameters of the grid simulator

Parameter/Size	Small	Large
Hosts (init., max, min)	32, 37, 27	256, 264, 248
Resource cap. (in MHz CPU)	N(5000, 875)	
Add host	N(625000, 93750)	N(437500, 65625)
Delete host	N(625000, 93750)	
Tasks (init., total)	384, 512	3072, 4096
Inter arrival	E(7812.5)	E(976.5625)
Workload	N(2500000000, 43750000)	
Security demands sd_i	U[0.6; 0.9]	
Trust levels tl_i	U[0.3; 1]	
Failure coefficient α	3	

We use the notation $U[x, y]$, $N(a, b)$ and $E(c, d)$ for uniform, Gaussian and exponential probability distributions respectively.

For activating the ANN module we divided the tasks and machines into 18 classes: 9 for processing power and trust level criteria (machines) and 9 for workload and security demand criteria (tasks). The ANN contains two hidden layers, the weight coefficients are in the range of $[-0.2; 0.2]$ and the learning rate is 0.01. The training set for ANN contains the characteristics of the tasks and machines and the task-machine matching results collected after the 500 runs of the simulator with inactive Neural Network module.

The key parameters for all types of genetic-based schedulers are presented in Table 3.

Table 3: GA-based schedulers settings

Parameter	Value
Evolution steps	5*n
Population size	60
Intermediate pop.	48
Cross probab.	0.9
Mutation probab.	0.15
Max time to spend	400 sec

We used the following three metrics to evaluate the scheduling performance:

- *Makespan*,
- *Flowtime*, and
- *FailureRate* F_r parameter defined as follows:

$$F_r = \frac{n_{failed}}{n} \cdot 100\%$$

where n_{failed} denotes the number of unfinished tasks, which must be rescheduled.

Each experiment was repeated 30 times under the same configuration of operators and parameters.

Results

In Fig. 1 and 2 we present the results of our experiments for four genetic-based risk resilient Schedulers.

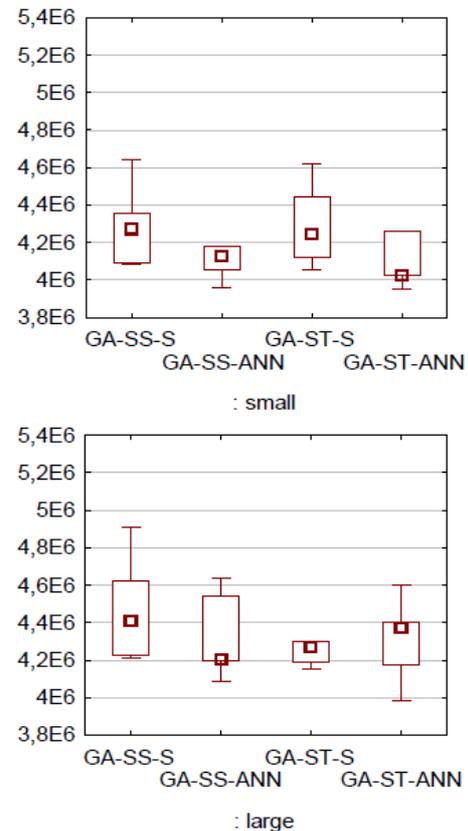


Fig. 1. The box-plot of the results for makespan.

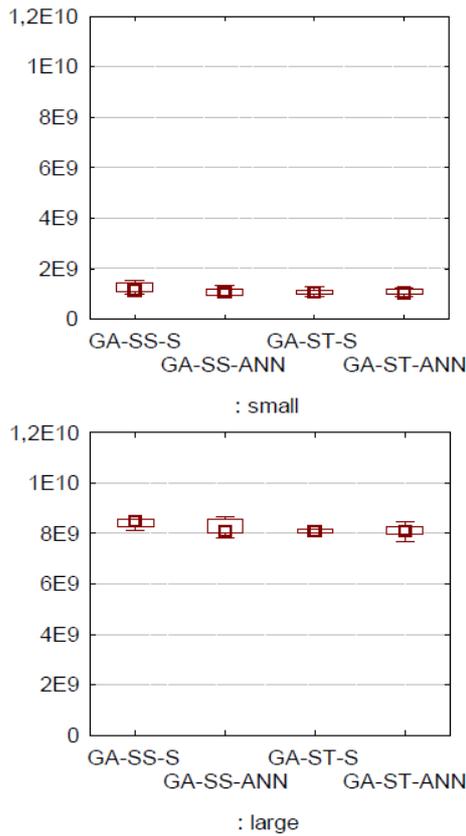


Fig. 2. The box-plot of the results for flowtime.

The best results in the makespan optimization have been achieved by both *GA-XX-ANN* schedulers. The efficiency of the ANN support can be observed especially in the ‘Large’ grid scenario. While in the small grid case the differences in the averaged makespan values are not so big, in second scenario *GA-SS-S* and *GA-ST-S* meta-heuristics significantly lag behind the secure schedulers. It can be also observed that in all instances the distribution of the results is asymmetric.

In the case of the flowtime minimization both *GA-XX-ANN* meta-heuristics outperform again the rest of the methods, however the differences in the results are not as significant as it was in the makespan case. Additionally, we noted that as the instance size is doubled, the flowtime values increase considerably for all applied schedulers, while the makespan is almost at the same level. Another observation is that all schedulers are rather ‘symmetric’ in sense of distribution of the results. The best relative effectiveness of the ANN support may be observed in ‘Large’ grid cases.

As the stopping criterion we consider in this work the maximal and a priori defined number of generations for which the GAs are activated. However, the solutions generated by ANN may not be improved by the GA metaheuristics, and the search process can be stopped because of the stagnation in the improvement of the solutions’ qualities. Therefore we additionally

‘measured’ for each scheduler the minimal time (expressed in the genetic epochs (generations)) necessary for the generation of the best solutions found by such a scheduler. The results are presented in Table 4. In parentheses we displayed the relative effectiveness parameter ef , which is calculated as the ratio of the minimal number of genetic epochs necessary for finding the optimal solutions and the stopping criterion, which is $5 \cdot n$, where n denotes the number of tasks in the system.

It can be noted that ANN module in most of the instances reduced the time necessary for finding the best possible solutions by each scheduler, so it may be also helpful in optimizing the whole process by stopping the algorithms much sooner than it was originally set, in most of the cases the time may be reduced approximately by 30–40 %.

Table 4: The number of genetic epochs necessary for the generation of the best solutions (efficiency parameter ef [%])

Strategy/Size	Small	Large
GA-SS-S	1831 (71.52%)	19002 (92.78%)
GA-SS-ANN	1622 (63.60%)	17830 (87.06%)
GA-ST-S	1703 (68.52%)	17993 (88.63%)
GA-ST-ANN	1511 (60.44%)	17910 (87.83%)

The effectiveness of the ANN support is confirmed by the lowest failure rates achieved by the *GA-XX-ANN* schedulers. The results for all four schedulers are presented in Table 5. In all instances the schedulers with the active ANN module outperform the other methods. The suboptimal solutions produced by ANN allow reducing the machine failures by 1% – 6%.

Table 5: Average values of Failure Rate parameter

Strategy/Size	Small	Large
GA-SS-S	6.104	8.943
GA-SS-ANN	4.88	5.026
GA-ST-S	9.218	5.744
GA-ST-ANN	4.093	7.894

CONCLUSIONS

In this paper we present the implementation of the Artificial Neural Network (ANN) as the support mechanism for risk resilient genetic-based schedulers in computational grids. Making a prior analysis of trust levels of the resources and security demand parameters of tasks, the neural network monitors the scheduling and task execution processes. The network learns patterns in input (initial tasks and machines characteristics) and produce the tasks-machines mapping suggestions as the outputs based on the stored data, which includes information about the resource failures. Then, based on the ANN ‘suggestions’ sub-optimal schedules are generated, which are then used to modify the

initialization procedures of genetic scheduling algorithms.

We have evaluated the proposed model under the heterogeneity, scalability and dynamics conditions using the Secure HyperSim-G Grid Simulator. We integrated a *Neural Network* module with the simulator, where *Minimal Completion Time (MCT)* algorithm is used for the sub-optimal schedules generation. The relative performance of four variants of the GA-based schedulers was measured for the makespan, flowtime and machines' failure rates metrics in secure scenario. We have demonstrated the efficiency of the schedulers supported by the ANN module in a fast reduction of the makespan and flowtime values and the improvement of the reliability of the resources. The best effectiveness of the ANN support can be observed in the makespan minimization in all considered grid scenarios. The obtained simulation results suggest that it is more resilient in the grid environment to pay some additional scheduling 'cost' due to verification of the security conditions instead of taking a risk on unreliable resources allocated.

Worth considering the extension of the comparison analysis to a wider set of meta-heuristic schedulers and introduce the multi-class tasks classification for generating the ANN outputs. It also will be interesting to provide some experimental study on the online scheduling, where some online learning mechanisms, like neuro-fuzzy systems (Nauck 1997) can be implemented to improve the scheduling results.

REFERENCES

- Abraham, A.; R. Buyya and B. Nath. 2000. "Natures heuristics for scheduling jobs on computational grids", *Proceedings of the 8th IEEE ACC*, India.
- Aguirre H.E., and K. Tanaka. 2007. "Working principles, behavior, and performance of MOEAs on MNK-landscapes", *European Journal of Operational Research*, vol. 181, 1670-1690.
- Ali, S.; H.J. Siegel; M. Maheswaran and D. Hensgen. 2000. "Task execution time modelling for heterogeneous computing systems", *Proceedings of Heterogeneous Computing Workshop*, 185-199.
- Bartschi Wall, M. 1996. "A Genetic Algorithm for Resource-Constrained Scheduling", *PhD Thesis*, Massachusetts Institute of Technology, MA.
- Garg, S.K.; R. Buyya and H.J. Segel. 2009. "Scheduling Parallel Applications on Utility Grids: Time and Cost Trade-off Management", *In Proc. of the 32nd ACSC*, Wellington, New Zealand. CRPIT, 91. Mans, B., Ed. ACS., 139-147.
- Grueninger, T. 1997. "Multimodal optimization using genetic algorithms", *Technical report*, Department of Mechanical Engineering, MIT, Cambridge, MA.
- Haykin, S. 1999. *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.
- Hotovy, S. 1996. "Workload evolution on the Cornell Theory Center IBM SP2", in *Job Scheduling Strategies for Parallel Proc. Workshop, IPPS'96*, 27-40.
- Kołodziej, J. and F. Xhafa. 2011. "Enhancing the genetic-based scheduling in computational Grids by a structured hierarchical population", *Future Generation Computer Systems*, vol. 27, DOI:10.1016/j.future.2011.04.011, 1035-1046.
- Nauck, D.; F. Klawonn and R. Kruse. 1997. *Neuro-Fuzzy Systems*, John Wiley & Sons.
- Song, S.; K. Hwang and Y.K. Kwok. 2005. "Trusted Grid Computing with Security Binding and Trust Integration", *Journal of Grid Computing*.
- Xhafa, F.; L. Barolli and A. Duresi. 2007. "Batch Mode Schedulers for Grid Systems", *International Journal of Web and Grid Services* 3(1), 19-37.
- Xhafa, F.; J. Carretero and A. Abraham. 2007. "Genetic Algorithm Based Schedulers for Grid Computing Systems", *International Journal of Innovative Computing, Information and Control*, vol. 3, No. 5, 1053-1071.

AUTHOR BIOGRAPHIES



DANIEL GRZONKA received his B.Sc. and M.Sc. degrees with distinctions in Computer Science at Cracow University of Technology, Poland, in 2012 and 2013, respectively. Actually, he is Ph.D. student at Jagiellonian University in cooperation with Polish Academy of Sciences. He is also a member of Polish Information Processing Society. His e-mail address is: grzonka.daniel@gmail.com.



JOANNA KOŁODZIEJ graduated in Mathematics from the Jagiellonian University in Cracow in 1992, where she also received the PhD in Computer Science in 2004. She works as an assistant professor at Cracow University of Technology. She has served and is currently serving as PC Co-Chair, General Co-Chair and IPC member of several international conferences and workshops including PPSN 2010, ECMS 2011, CISIS 2011, 3PGCIC 2011, CISSE 2006, CEC 2008, IACS 2008-2009, ICAART 2009-2010. Dr Koodziej is a EB member and guest editor of several peer-reviewed international journals and author and co-author of many publications in high quality peer reviewed international journals. Her e-mail address is: jokolodziej@pk.edu.pl and her Web-page can be found at <http://www.joannakolodziej.org>.



JIE TAO got her PhD degree at the department of Computer Science of Munich University of Technology, Germany. She is currently a senior research associate at the Steinbuch Centre for Computing, Karlsruhe Institute of Technology. Dr. Tao's research work targets mainly on parallel and distributed computing, data-intensive computing as well as Grid and Cloud computing. Dr. Tao has published a number of articles in peer-reviewed international journals and leading conferences. She serves as co-chair or PC member in a set of international conferences and workshops. She is a guest editor of several international journals. Her e-mail address is: jie.tao@kit.edu.

Hybrid Architecture for Simulation of Blood Flow with Foreign Bodies

Łukasz Faber, Krzysztof Boryczko, Marek Kisiel-Dorohinicki
AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: {faber,boryczko,doroh}@agh.edu.pl

KEYWORDS

agent-based simulation, dissipative particle dynamics

ABSTRACT

The new methods of diagnosis are often more sensitive and speed up its process allowing for a more effective treatment. Yet they need to be carefully tested before they can be used in practice. The paper concerns the problems of developing nanorobots, which would circulate in the bloodstream of the human body gathering information about its condition. Simulation of these new ways of diagnosis is the main motivation behind the presented work. Introduction of such nanorobots into the bloodstream puts several requirements on their mechanical properties. Thus, our first goal is to build a model to simulate the blood flow with nanorobots present in the capillary vessels and medium-sized blood vessels. We split it into two separate initial models and implementations. The first one is a simulation of the blood flow using particle-based methods in order to determine the appropriate mechanical parameters of nanorobots and verify their behavior. The second one is a multi-agent simulation that will allow to evaluate the usefulness of the data collected and prototype the system performing functions of nanorobots within the specified constraints.

INTRODUCTION

Early and accurate diagnosis is one of the most important factors that make it possible to provide a patient with an effective treatment. This leads to constant improvements to traditional and popular methods of diagnosis – usually noninvasive (e.g., diagnostic imaging). However, technological development makes it possible to introduce new, better, more sensitive methods that speed up the process of diagnosis (e.g., using the optoelectronics for marker detection in exhaled air). Moreover, it is not an end of possible ways to perform an early diagnosis. The next step is nanotechnology and particularly nanorobotics. In our work, we consider the existence of nanorobots able to circulate along with blood cells in the circulation system of the human body. These nanorobots would have the ability to gather and transmit data about the (internal) state of the body. On the basis of this data it would be possible to identify a possible condition of a patient.

The goal of the paper is to elaborate on possible techniques and tools for simulation of nanorobots circulating in the vessels of the human body and gathering information about its state. We try to identify the best solutions covering different aspects of the simulation and propose an architecture with their

cooperation in mind. Simulation of blood flow and capillary vessels is usually performed using particle-based methods [1]. However, it is usually difficult to extend these models with more ‘intelligent’ behavior. When using the term ‘*intelligent behavior*’ we think of some specific aspects: 1) generation of the biological properties of the vessels and blood (e.g., temperature, permeation of other substances); 2) behavior of injected nanorobots, and 3) external components of the system (e.g., external sensors or data collectors). On the other hand, agent-based simulations are perfect for such work. Therefore, we want to add agent-based simulations to the particle models.

From this background we can separate two aspects of our work:

- 1) Simulation of the blood flow in capillary vessels of the human body with the additional presence of foreign bodies using particle-based methods.
- 2) Estimation of the scope and usefulness of the data obtainable from such an environment using a distributed multi-agent system.

These aspects, although they touch completely different fields, are strictly related. The former will help to specify *physical* parameters and constraints. The latter will possibly lead to extending the particle simulation in order to emulate the environment properties in a more realistic way.

In order to reach these goals we are building two simulation applications:

- 1) one based on particle models (Dissipative Particle Dynamics – DPD, Smoothed Particle Hydrodynamics – SPH, Smoothed DPD – SDPD, TC-FPM – Thermodynamically Consistent Fluid Particle Model) for simulating the blood flow with nanorobots present,
- 2) one agent-based that will be able to simulate behaviors of nanorobots and their environment for testing the data collection possibilities; this application will be also responsible for extending the previous one with biological data.

Both of these applications require construction of the model, software implementation, testing and running large simulations.

For the first application, we built the model of the blood flow in capillaries, based on modern particle methods: SDPD and TC-FPM. The software is built using the C language. In the later stages the OpenCL reimplementations are planned.

Similarly, for the second component, we developed the agent model for the simulation of nanorobots as a preliminary

requirement for the implementation of the simulation platform. We should note, that the characteristics of these models are quite specific for the problem: the agents are very small but, on the other hand, quite capable in terms of their functions. They can perform full range of complicated behavior: sensing, acting, communication, movements. To the best of authors knowledge, there are no similar, ready-to-use models.

Apart from discussing the techniques for both simulation models, in the paper we present the architectural description of both components and propose the mechanism of their interaction. The actual evaluation of the architecture is based on the experiences gained during independent runs of the simulations of the blood flow and the agents, yet the results encourage further work in this direction.

BACKGROUND AND TECHNICAL SOLUTIONS

In this section we will review both discussed simulation models and most popular tools used with them.

Blood Flow Simulations

Blood is a highly complex tissue with its cells suspended in a liquid environment known as blood plasma. An important fact is that the blood and the vascular bed are the integral parts of all tissues of the body. Thus, disorders in physiology of individual body organs result in changes to the composition of the blood and vice versa – blood disorders and changes in its composition have a significant effect on the entire body.

The physiological task of the blood is transport. The blood transports oxygen, carbon dioxide, nutrients, hormones, metabolic products and cells. The task of the oxygen carrier is fulfilled by erythrocytes which are the most numerous blood cells and constitute about 45% of its volume. Their shape resembles a biconcave lens with a diameter of about 8 μm . This shape ensures a relatively high surface area to volume ratio and allows to adjust size to the capillaries. Research has shown that erythrocytes are able to flow through the vessels with the diameter of about 4 μm .

Number, shape and flexibility of erythrocytes have a direct impact on macroscopic parameters describing the hydrodynamic properties of blood. They have been a subject of numerous studies and simulations. A lot of models were developed in order to study the flow of blood in large blood vessels as well as the behavior of its cells when traveling through the capillaries. It is worth to note that there is a need to adjust the time-space scale of the used simulation method to the simulated phenomenon. Thus, in flow simulations in large (macroscopic) blood vessels the CFD methods or the Smoothed Particle Hydrodynamics (SPH) method are used. Simulation of the blood flow in capillaries, in mesoscale, was for a long time quite troublesome as there was no adequate model for this scale. First models were relatively large simplifications of the phenomenon [2], [3], [4]. Only in 1992, the Dissipative Particle Dynamics (DPD) method was proposed [5]. It was designed to simulate phenomena happening in the mesoscopic scale. This method had several artifacts that were later removed in a method called Fluid Particle Model (FPM). As shown in [6] these methods and their combinations are well-suited for the accurate modeling of complex fluid flows in the size range 10 nm to 100 μm .

On this basis, the model of blood flow in the capillaries was proposed that took into account the flexibility of erythrocytes, their interactions with the walls of the capillaries and the presence of fibrinogen [7], [8].

Multi-Agent Systems

In a Multi-Agent System (MAS), the process of solving a computational problem is decomposed into autonomous, intelligent entities called agents. Agents can actively follow some goals, based on their beliefs (perceived environment), plan their actions ahead and learn from experiences.

In order to achieve this, MAS are often used along with other methods from Artificial Intelligence or Machine Learning. Agents may also own resources and interact with each other. The structure of the system is expected to emerge from these decentralized agent interactions. MAS are thus a bottom-up and decentralized approach to system design and problem solving.

Multi-agent systems and, more generally, the concept of an intelligent agent have found multiple applications, both as a way to model complex systems and as a programming paradigm to implement them. Several established agent-based technological solutions exist, including FIPA compliant, fully-fledged environments like JADE or JADEX, where agents are a basic unit of software abstraction. As an example of another approach, there are also minimalistic and easy-to-use tools like NetLogo, where agents are only present at a domain level, as means of problem decomposition.

The first class of systems can be used to solve any problem that benefits from using an agent-oriented approach. However, in some particular classes of applications, this can be very inefficient. This happens especially in the case of simulation and computational applications, where agents and their behaviors are well defined. Such MAS might not need to be open to other systems, require the possibility of code migration nor support FIPA-compliant communication.

In these cases, systems of the second class are more efficient and a much better choice. However, they suffer from other drawbacks, as they do not support component-oriented approach, which affects code reusability and make more complex problems hard to program. Also, these systems are usually not suited for running in distributed environments. In other words, they do not scale well with bigger problems or more complex systems.

Usually, Multi-Agent Systems have following properties [9]: agents are autonomous and independent, agents are aware only of their local environment; in other words, no agent has all the data in the system, there is no central control component (i.e., an agent that would be able to supervise the whole system). Some researchers add also asynchronicity of the system to this set of properties [10]. Moreover, to execute their tasks in such a system agents usually need to have means of communication and mutual interactions. The way they are handled depends on the environment and system requirements.

Agent-based (and multi-agent) systems are well-researched and extensively used solutions to a very wide range of different problems: transportation, metaheuristic problem solving [11], multi-robotics, social simulations, etc. Their properties make

them well-suited for distributed systems [12]: they offer modularity and scale efficiently.

The question arises: why did we choose multi-agent systems? The simple answer is: agents are successfully used in multi-robotics scenarios due to their properties [13].

The MAS properties, that we described above, correspond clearly with the properties of the nanorobotic system that could be located in the blood vessels. In such conditions: nanorobots need to be autonomous as there is no possibility for all of them to communicate, nanorobots are unable to ‘know’ the whole environment as they are constrained in terms of computational power and storage capabilities, there is no possibility to introduce one central component that would be able to supervise all nanorobots. Moreover, the system is completely asynchronous. It also needs to adapt to constantly changing properties in terms of agents relative and absolute spatial positions, their communication and interaction partners.

As a second phase requires a simulation of a patient body properties, we will need to use an existing or implement a new agent-based simulation platform. For this purpose we should briefly review some of the possible choices of simulation platforms.

There exists a plethora of multi-agent frameworks which may be used to support the construction of simulation systems. Some of them are oriented to specific kinds of simulation (see [14], [15]): e.g., simulating of movement of entities with 3D visualization (e.g., breve [16]), possibility of visual programming (e.g., SeSam [17]). When looking for universal agent-based simulation frameworks (especially in open-source software domain), one should consider such products as, for example, Galatea [18], RePast [19], Mason [20]. Other ones are general-purpose agent-based programming frameworks (e.g., JADE [21]) that may be of course adapted to any kind of simulation. MadKit [22] should also be mentioned as a framework for simulating complex populations (following Agent/Group/Role paradigm).

The frameworks described in the next paragraphs were selected as the most promising examples of general-purpose agent-based simulation frameworks in the open-source market.

MASON is an agent-oriented simulation framework developed at George Mason University. It is advertised as fast, portable, 100% Java based. The multi-layer architecture brings complete independence of the simulation logic from visualization tools. The models are self-contained and may be included in other Java-based programs.

The programming model of MASON follows the basic principles of the object-oriented design. An agent is instantiated as an object of a class, added to a scheduler and its *step* method is called during the simulation. There are no predefined communication nor organization mechanism. There are neither ready-to-use distributed computing facilities nor component-oriented solutions.

RePast is a widely used agent-based modeling and simulation tool. It has multiple implementations in several languages and built-in adaptive features such as genetic algorithms and regression. The framework uses fully concurrent discrete event scheduling. Dynamic access to the models in the runtime (introspection) is possible using a graphical user interface.

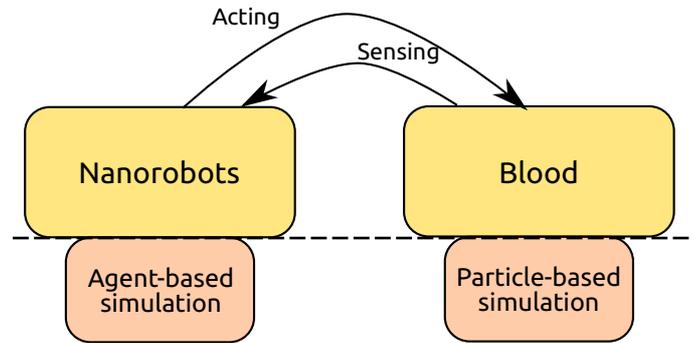


Fig. 1: Overview of the architecture. We can pinpoint two layers: conceptual (upper) and implementational (lower). The communication in the former is related to the way agents obtain and generate information. Agents sense the environment and (possibly) act on it.

In Repast Symphony, a new organizational concept called ‘context’ was introduced. It consists of a group of unorganized agents (they may be organized using so-called projections) and may create a hierarchical structure (context can have many sub-contexts and so on). This idea affects the perception of agents in such way, that an agent in the sub-context also exists in the parent context, but the reverse is usually not true.

MadKit is a modular and scalable multi-agent platform written in Java, aimed at modeling different agent organizations, groups and roles in artificial societies. It is built based on a so called Agent/Group/Role organizational model, using a plugin-based architecture. The architecture of MadKit is based on micro-kernels which provide only the basic facilities: local messaging, management of groups and roles, launching and killing of agents. Other features (remote messages, visualization, monitoring and control of agents) are performed by agents. Both thread based and scheduled agents may be developed.

Simulations do not require any particular structure or model to run. However, it is possible to add an arbitrary scheduler or create complex agent communities and relationships. Agents can locate other agents having some specific role or belonging to some particular group. Agents can communicate with each other using these roles or group membership (i.e. using broadcast messages).

ARCHITECTURE

The system has two heterogeneous components that can be discussed independently. We have built two separate models for initial tests. The basic idea behind the system is to use particle-based simulation techniques for the environment and agent-based ones for the foreign bodies ‘injected’ into the flow. In our particular case we define environment as capillary vessels, and the external bodies are nanorobots. An overview of the architecture is shown in Figure 1.

The system is parallel by the nature of the performed computations. The current implementations are based on C (Message Passing Interface – MPI) and Java, but future implementations will be prepared mostly for highly-parallel environments like GPGPU.

Particle-based Component

For the first phase, the model of the blood flow in capillary vessels will be built. It will be based on modern particle methods that take into consideration internal state of particles (thermodynamically consistent) – SDPD (Smoothed DPD) [23] and TC-FPM (Thermodynamically Consistent FPM) [24]. Simulations will make it possible to set the mechanical parameters of the introduced objects that are required for their proper functioning in the circulatory system. Moreover, the physical parameters of the environment that should be possible to determine around the flowing object will be defined. Application of these simulation methods will allow to determine more parameters, including the internal temperature of tissues as its differences, in the context of other information, may indicate the presence of cancer.

During the first phase we will implement the simulation application. Then, we will perform extensive tests in order to obtain proper parameters of the simulation and to ensure high reliability and optimization (in terms of simulation speed). As mentioned earlier, we plan to run the simulation in a highly parallel environment. The obtained parameters will be used for specifying constraints of an agent operating ‘inside’ a nanorobot.

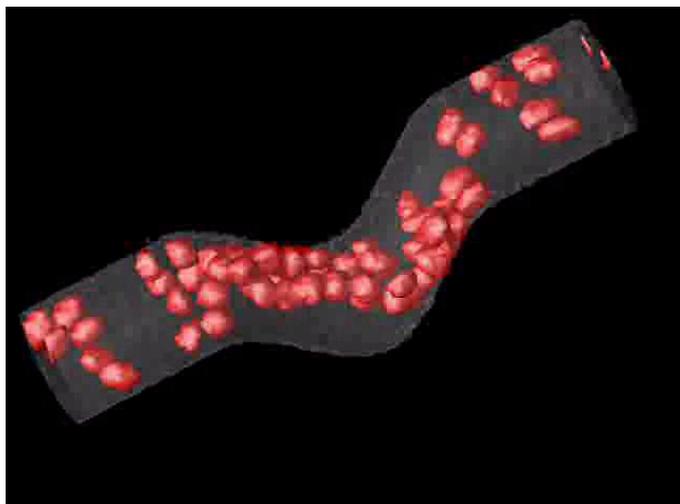


Fig. 2: Flow of the bodies in a larger vessel.

Figures 2 and 3 show a flow of the bodies in larger vessel during simulations with aim to determine how their shapes influence viscosity. Figure 4 is the part of determining mechanical properties of the bodies. We try minimize their influence on the flow of blood. In our simulations we use properties similar to erythrocytes but we omit proportions of the volume to the area as it is relevant only for an oxygen transport.

Our goal is to specify shape of a nanorobot. Some of the sample, initial results include the computation of Reynolds numbers for flat and biconcave bodies in relation to the number of simulation steps. As mentioned before, we do not have to keep proportions that are normally required for erythrocytes because there is no oxygen transport involved. If possible, we would prefer to increase the volume but keep all of the mechanical properties. Thus, our first attempt is to use not

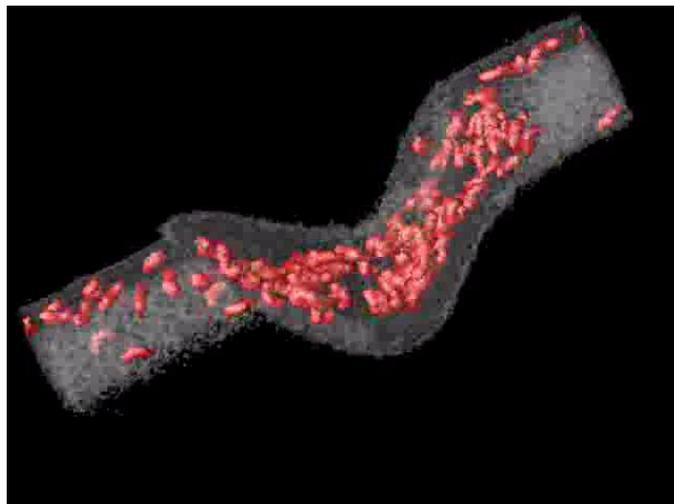


Fig. 3: Flow of the bodies in a larger vessel.

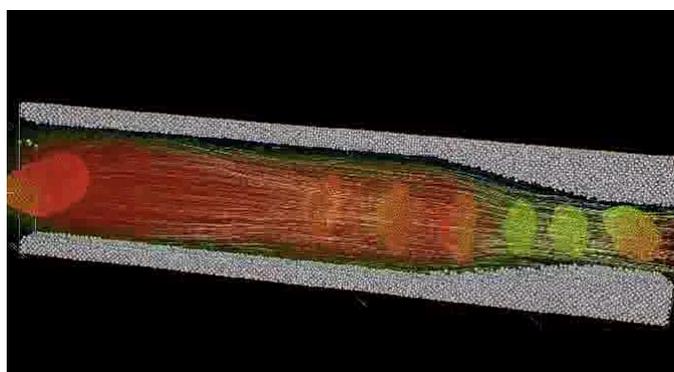


Fig. 4: Visualization of the flow in simulations leading to determining of mechanical properties of the bodies.

biconcave but flat nanorobots. We compared the behavior of these two possibilities (flat and biconcave nanorobots). Figures 6, 7 and 8 show shapes of the nanorobots successively in: 0th, 2500th and 4000th steps of the simulation. It is noticeable that flat ones are less flexible and this reflects on the blood flow. This fact is also visible in Figure 5.

Agent-based Component

The agent-based component should have following general properties:

- It should consist mostly of small agents strictly constrained in terms of computational capacity and the available memory size.
- It should be able to simulate agents separation in human body and spatial distribution.
- It should be possible to introduce larger entities into the system (but external to the body) that will be able to execute more complicated tasks (like data gathering).
- There is a need to introduce spatial properties (positions) and neighborhood for both the agents playing the role of the environment and those controlling nanorobots. For the former, because they are parts of a specific spatial

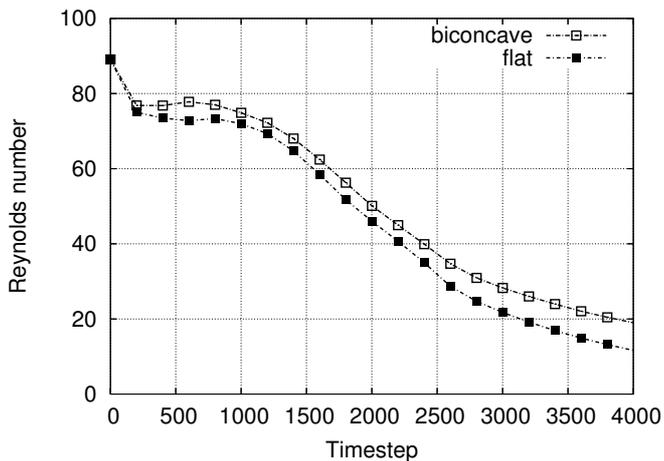


Fig. 5: Value of the Reynolds number (in program-dependent units) in consecutive steps of the simulation.

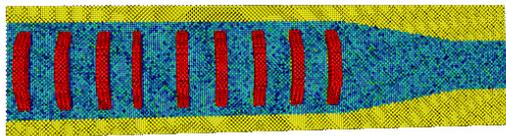


Fig. 6: Visualization of the 0th step of the simulation (initial configuration).

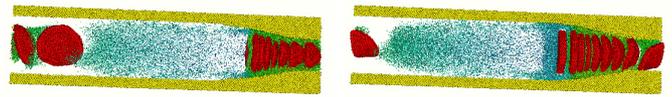
structure, and for the latter because there is a need to know which agents can communicate in order to simulate aforementioned communication problems.

- The nanorobots need to pass the collected data to an external component that will be able to process it having a more broad view (as it will have more information about the whole system and thus about the body).

There are several possibilities how to implement such an agent-based simulation platform. Initially, we test our requirements using the AgE platform [25]. The AgE platform includes two types of agents: heavyweight agents and lightweight ones.

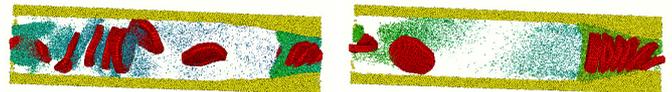
Heavy agents are realized as separate threads, as in the JADE platform. They communicate through asynchronous message passing. This is an effective model when the number of agents is small or agent interactions are sparse (i.e. each agent communicates only with a small number of other agents).

Moreover, AgE introduces lightweight agents. Populations of such agents are thread-contained and execute pseudo-concurrently, one step each at a time. Some restrictions on how agents may change each other state, along with communication constraints, described further below, emulate concurrency effects and execution interleaving. From these agents perspective, they are effectively processed in parallel.



(a) Biconcave nanorobots. (b) Flat nanorobots.

Fig. 7: Visualization of the 2500th step of the simulation.



(a) Biconcave nanorobots. (b) Flat nanorobots.

Fig. 8: Visualization of the 4000th step of the simulation.

Thus, the platform API available to heavyweight and lightweight agents is very similar. The main difference is in efficiency and determinism, as needed in a particular case.

Each agent in AgE belongs to some environment. Agents can communicate with each other within their environment or query it to acquire some information. In the case of heavy agents, the environment simply consists of multiple distributed nodes. All heavy agents on all connected nodes belongs to it. When it comes to lightweight agents, environments are treated as agents themselves. These *aggregate* lightweight agents also have properties, behavior, and also execute in some higher level environment. Thus lightweight agents form hierarchies, as shown in Figure 9. The root agent is a heavy one and is called a *workplace*. A workplace encompasses all the hierarchy within its thread and initiates step based, pseudo-concurrent execution.

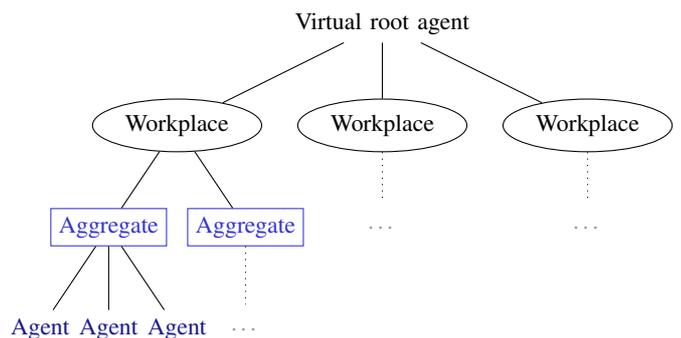


Fig. 9: Agent tree structure in AgE. Workplaces are units of distribution among separate AgE nodes. Each workplace may consist of any number of aggregates and simple agents. The *virtual root agent* represents a common root for all workplaces which handles communication facilities on their level.

In future, additional functionalities of the platform should include monitoring of the system for purpose of having an online information about its state and distribution of the simulation.

At least three broad kinds of agents are introduced:

- 1) one for simulation of the environment,
- 2) one for simulation of the nanorobots, and
- 3) one for being in role of possible external components.

Their general relations are shown in Figure 10. The red circles are agents that act as body tissues. They are spatially close to each other and can communicate locally in a peer-to-peer fashion. The blue circles are external component agents that also are spatially close and have efficient and a reliable peer-to-peer or leveled communication. The yellow circles are the agents controlling nanorobots. They can communicate with each other when they are in close neighborhood. They can also communicate with the external components (the bottom part of the figure) or with the body tissues (the upper part of the figure). Obviously, the interaction with the body tissues is regarded as ‘communication’ only from the simulation point of view. In the model such interactions are analogous to ‘perceiving of’ and ‘acting on’ the environment.

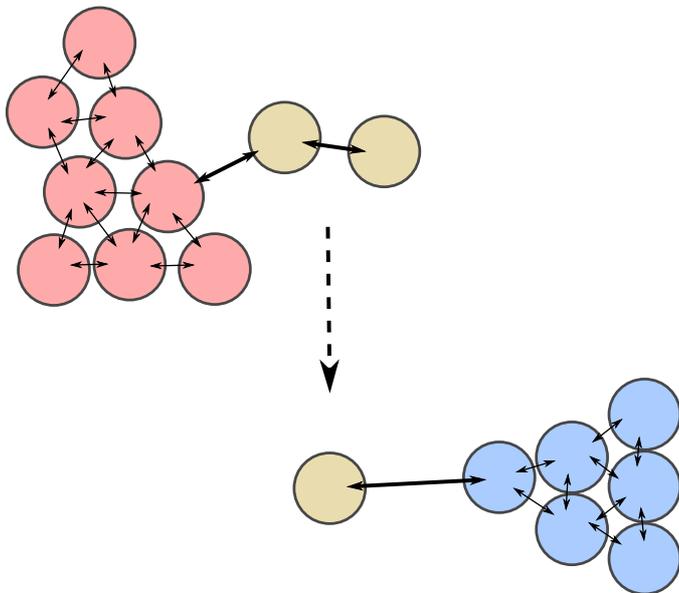


Fig. 10: Communication between agent types.

Communication between both models may be realized in many ways. Our first, the simplest, attempt was to perform simulations separately (we call it *offline* version). We run the particle-based simulation and we obtain physical properties of the system which we then submit to the agent-based simulation. Data acquisition can be done in two ways: we can dump all simulation data in every step or we can interpolate physical properties (like temperature, speed, etc.) during the run and dump only these values.

In order to allow feedback from the agent-based simulation to the particle-based one, we need to implement concurrently running simulations that will be exchanging data *online*. In such scenario, the blood flow simulation will be forwarding

data (full or selected interpolated properties) to the agent system in every step. The second system will be performing its own computations and will return results to the first one. As the complexity and execution time of the particle-based simulation is higher than that of the agent-based simulation, both systems will be able to run simultaneously. Sample realization is shown in Figure 11.

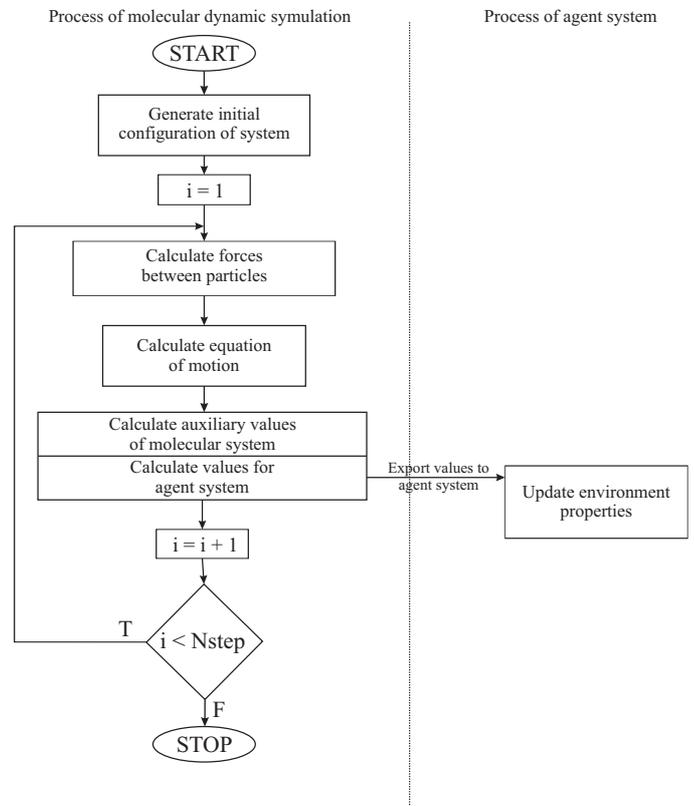


Fig. 11: Processing of particle simulation with data forwarding to the agent system.

CONCLUSIONS AND FUTURE WORK

We have presented the initial model and implementation of the hybrid simulation system for blood flows with foreign bodies. The system comprises of two heterogeneous simulation: the particle-based one and the agent-based one. The former is used as a tool to simulate physical properties of the environment (i.e. capillary vessels) whilst the latter is responsible for the generation of the additional, biological properties of the environment.

Achievement of the aforementioned objectives opens up several possibilities for further research in all of related areas. For example, for agent systems it could be improving the behavior of the nanorobots, diagnosis deduction, etc. In long-term it should also lead to creation of working prototypes.

The architecture was described in the context of the work on the system for obtaining data about nanorobots in the bloodstream of the human body.

The most important aspect of future work is a merge of these two, currently separate models and applications into

one, heterogeneous model possibly working as one application or data-exchanging cluster of applications. Moreover, our implementation goals are related to creating high-performance simulations using GPGPU.

ACKNOWLEDGEMENT

The research reported in the paper was supported by the grant “Hybrid model of the early detection of internal diseases based on the paradigm of interacting particles and multi-agent system” (No. UMO-2013/09/N/ST6/01011) from the Polish National Science Centre.

REFERENCES

- [1] W. Dzwinel, K. Boryczko, and D. A. Yuen, “A discrete-particle model of blood dynamics in capillary vessels,” *Journal of colloid and interface science*, vol. 258, no. 1, pp. 163–173, 2003.
- [2] R. Hsu and T. Secomb, “Motion of nonaxisymmetric red blood cells in cylindrical capillaries,” *Journal of biomechanical engineering*, vol. 111, no. 2, pp. 147–151, 1989.
- [3] S. R. Keller and R. Skalak, “Motion of a tank-treading ellipsoidal particle in a shear flow,” *Journal of Fluid Mechanics*, vol. 120, pp. 27–47, 1982.
- [4] P. Mazon, S. Muller, and H. El Azouzi, “Deformation of erythrocytes under shear: a small-angle light scattering study,” *Biorheology*, vol. 34, no. 2, pp. 99–110, 1997.
- [5] P. Hoogerbrugge and J. Koelman, “Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics,” *EPL (Europhysics Letters)*, vol. 19, no. 3, p. 155, 1992.
- [6] K. Boryczko, W. Dzwinel, and D. A. Yuen, “Parallel implementation of the fluid particle model for simulating complex fluids in the mesoscale,” *Concurrency and computation: practice and experience*, vol. 14, no. 2, pp. 137–161, 2002.
- [7] —, “Dynamical clustering of red blood cells in capillary vessels,” *Journal of Molecular Modeling*, vol. 9, no. 1, pp. 16–33, 2003.
- [8] —, “Modeling fibrin aggregation in blood flow with discrete-particles,” *Computer methods and programs in biomedicine*, vol. 75, no. 3, pp. 181–194, 2004.
- [9] M. Wooldridge, *An introduction to multiagent systems*. John Wiley & Sons, 2009.
- [10] R. A. Flores-Mendez, “Towards a standardization of multi-agent system framework,” *Crossroads*, vol. 5, no. 4, pp. 18–24, 1999.
- [11] A. Byrski and M. Kisiel-Dorohinicki, “Agent-based meta-heuristic approach to discrete optimization,” in *Complex, Intelligent and Software Intensive Systems (CISIS), 2011 International Conference on*. IEEE, 2011, pp. 508–512.
- [12] H. S. Nwana, “Software agents: An overview,” *The knowledge engineering review*, vol. 11, no. 03, pp. 205–244, 1996.
- [13] A. Soriano, E. J. Bernabeu, A. Valera, and M. Vallés, “Multi-agent systems platform for mobile robots collision avoidance,” in *Advances on Practical Applications of Agents and Multi-Agent Systems*. Springer, 2013, pp. 320–323.
- [14] C. Nikolai and G. Madey, “Tools of the trade: A survey of various agent based modeling platforms,” *Journal of Artificial Societies & Social Simulation*, vol. 12, no. 2, 2009.
- [15] S. F. Railsback, S. L. Lytinen, and S. K. Jackson, “Agent-based simulation platforms: Review and development recommendations,” *Simulation*, vol. 82, no. 9, pp. 609–623, 2006.
- [16] J. Klein, “Breve: a 3d environment for the simulation of decentralized systems and artificial life,” in *Proceedings of the eighth international conference on Artificial life*, 2003, pp. 329–334.
- [17] N. Ventroux, A. Guerre, T. Sassolas, L. Moutaoukil, G. Blanc, C. Bechara, and R. David, “Sesam: An mpoc simulation environment for dynamic application processing,” in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 1880–1886.
- [18] J. Dávila and M. Uzcátegui, “Galatea: A multi-agent simulation platform,” in *Proceedings of the International Conference on Modeling, Simulation and Neural Networks*, 2000.
- [19] N. Collier and M. North, “Parallel agent-based simulation with repast for high performance computing,” *Simulation*, vol. 89, no. 10, pp. 1215–1235, 2013.
- [20] S. Luke, C. Cioffi-Revilla, L. Panait, and K. Sullivan, “Mason: A new multi-agent simulation toolkit,” in *Proceedings of the 2004 SwarmFest Workshop*, vol. 8, 2004.
- [21] F. Bellifemine, A. Poggi, and G. Rimassa, “Jade—a fipa-compliant agent framework,” in *Proceedings of PAAM*, vol. 99, no. 97-108. London, 1999, p. 33.
- [22] O. Gutknecht and J. Ferber, “The madkit agent platform architecture,” in *Infrastructure for Agents, Multi-Agent Systems, and Scalable Multi-Agent Systems*. Springer, 2001, pp. 48–55.
- [23] P. Espanol and M. Revenga, “Smoothed dissipative particle dynamics,” *PHYSICAL REVIEW-SERIES E-*, vol. 67, no. 2; PART 2, pp. 026 705–026 705, 2003.
- [24] M. Serrano and P. Espanol, “Thermodynamically consistent mesoscopic fluid particle model,” *Physical Review E*, vol. 64, no. 4, p. 046115, 2001.
- [25] Ł. Faber, K. Piętak, A. Byrski, and M. Kisiel-Dorohinicki, “Agent-based Simulation in AgE Framework,” in *Advances in Intelligent Modelling and Simulation*. Springer, 2012, pp. 55–83.

ABOUT THE AUTHORS

Łukasz Faber obtained his M.Sc. in 2012 at AGH University of Science and Technology in Cracow and is currently a Ph.D. student at the Department of Computer Science of AGH-UST. His research interests include agent-based modeling and distributed systems.

Krzysztof Boryczko received his Ph.D. degree in computer science in 1992 and D.Sc. degree in computer science in 2004, both from the AGH University of Science and Technology, where he is now Associate Professor at the Department of Computer Science, Faculty of Computer Science, Electronics and Telecommunications. His research interests focus on large-scale simulations with particle methods and on implementations of particle methods on Graphical Processing Units. He is also interested in scientific visualisation, feature extraction and clustering algorithms for analysis of simulation data.

Marek Kisiel-Dorohinicki obtained his Ph.D. in 2001 at AGH University of Science and Technology in Cracow. He works as an assistant professor at the Department of Computer Science of AGH-UST. His research focuses on intelligent software systems, particularly using agent technology and evolutionary algorithms, but also other soft computing techniques.

Realistic mobility simulator for smart traffic systems and applications

Cosmin-Stefan Stoica, Ciprian Dobre, Florin Pop

University Politehnica of Bucharest

Bucharest, Romania

cosmin.stoica@cti.pub.ro, ciprian.dobre@cti.pub.ro, florin.pop@cti.pub.ro

Abstract—Cars have become essential elements of modern life. But nowadays the increasing number of cars also leads to problems: pollution, traffic jams, wasted time spent in traffic because of traffic bottlenecks, etc.. Traffic cannot cope anymore with the rate of car usage today. Fortunately, in the last years various Intelligent Transportation Systems (ITS) demonstrate innovative services relating to different modes of transport and traffic management, and enable various users to be better informed and make safer, more coordinated, and 'smarter' use of transport networks. For such systems, a lot of attention has been dedicated to develop realistic models of roads and of the vehicle mobility, by exploiting the extensive literature developed in the field of transportation systems, e.g., models of how cars move along a road. Here, we present the realization of a more-realistic simulation tool, *Sim²Car*. The simulator uses as input data real-life traces collected over long periods of time, involving potentially thousands of drivers, to provide a realistic support mobility model for high-level methods and techniques designed to optimize traffic. On top, developers can simulate advanced traffic networking solutions and intelligent transportation applications, under real-world conditions. We present our results for an application designed to optimize the costs involved in using navigators, that actively incorporate solutions borrowed from opportunistic computing and context data to optimize the user's experience.

Index Terms—realistic simulator, traffic simulation, congestion, real-world mobility data, uncertain location data and correction algorithms

I. INTRODUCTION

Traffic congestion is happening in many urban environments. The road infrastructure capacities cannot cope with the rate of increase in the number of cars. This, coupled with traffic incidents, work zones, weather conditions, make traffic congestion a major concern for municipalities and research organizations [1]. Advanced traffic control technologies may lead to more efficient use of existing road network systems, resulting in reduced traffic congestion, delays, emissions, energy consumption and improved safety. The communication capabilities provided by modern wireless technologies and mobile devices offer opportunities for the development of traffic control applications where cars and devices within the road infrastructure collaborate to solve traffic problems. Thus, Intelligent Transportation Systems (ITS) have socio-economic advantages towards reducing traffic congestion, the high number of traffic road accidents, etc.. Coupled with advances brought by modern vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication technologies,

ITS research can lead to a plethora of traffic applications (e.g., collision avoidance, road obstacle warning, safety message dissemination, etc.), traffic information and infotainment services (e.g., games, etc.) [2].

Simulation has a major role in designing and evaluating such solutions. In the last years increase attention has been dedicated to developing realistic models of vehicle mobility (i.e., we previously proposed such a mobility model based on the social behavior of drivers in [3]). However, such models are still unable to accurately reproduce the way drivers behave in real-world cities. Several well-known simulators (ns-2, ns-3, SWANS, OMNET++, OPNET, etc.) are able to take as input trace files, but such files have to be generated by specialized vehicle mobility models (e.g., Vanet-MobiSim, SUMO), which mimic vehicle mobility behaviors. However, in some scenarios, the interdependencies among vehicle communication, or their mobility patterns, are still insufficiently close to real-world. For example, when a congestion/accident occurs, the vehicle mobility changes due to triggers from an ITS service.

In this paper, we propose to advance the field, by providing realistic simulation for the evaluation of ITS, using real-world mobility traces. The input is represented by real-world mobility data, collected over a time interval from traffic, in different cities (i.e., we previously developed such a traffic collector app for Android, that uses crowd-sensing techniques [2]). The few previous attempts to create realistic traffic models were mostly based on wrong conclusions [4], mostly because of uncertainty in GPS readings. We advance on that, and present the *Sim²Car* simulator. On a microscopic level, each car moves according to the mobility trace. On a microscopic level, we use OpenStreetMap (OSM) to build the street graph, and to correct the mobility traces using an algorithm to correct the GPS inaccurate position of each car.

We present experiments with mobility traces of 500 cabs from San Francisco Bay, C.A., and prove the capabilities of the simulator – in this case, *Sim²Car* is able to successfully cope with a simulation of realistic taxi movement, covering a total distance of trajectories of approx. 9 million kilometers.

The rest of the paper is structured as follows. Section II discusses related work. Section III presents the proposed solution, while Section IV describes the implementation details. Section V presents experimental results and analysis of the obtained performance. Section VII concludes and presents future work.

II. RELATED WORK

Different authors previously attempt to provide realistic conditions for the evaluation in simulation of various traffic and vehicular communication mechanisms. SUMO [5], [6] is one of the mostly cited purely microscopic open-source simulator for traffic, designed to handle large road networks. SUMO integrates a large suite of tools for street network processing and traffic patterns generation. Internally, SUMO uses a graph to model the network of modeled streets within a city. The streets (with one or more lanes) are represented as edges of an oriented graph, with nodes being the intersections [6]. Edges can keep additional information, such as speed restrictions, street category, traffic lights position, and other traffic signs, etc.. The netgenerate tool is capable in SUMO to automate the generation of streets networks (as grid-networks, circular spider networks and random networks). For realistic simulations, SUMO provides netconvert, a tool capable to transform digital world maps (e.g., in OpenStreetMap format) into SUMO-format.

MOVE [7], [8] extends SUMO with advanced usability capabilities. The tool automates the generation of simulation scenarios, and adds new features, such as the capability to generate traces from Tiger database. However, its current implementation does not offer support for traffic parameters (as compared to the standard SUMO generated traces, it does not include features as lane change or obstacle mobility model), and cannot easily support large simulations (with many cars).

VNSim [9], [10] is a microscopic simulator capable to generate synthetic, social-driven, mobility, with a network model on top. Unfortunately, even if the synthetic mobility model in VNSim can be generated starting from TIGET maps, it still lacks the realism needed to correctly evaluate ITS mechanisms and techniques.

Analyzing previous ITS simulators, they suffer from the problem of huge memory consumption in case of intensive simulation scenarios. We analyzed the possibility to use previous simulators with mobility traces with more than a thousand cars (a realistic assumption, considering real-world conditions, where most likely more than a thousand car will use the application under test), and they all fail to scale. For example, trying to avoid this memory problem, SUMO (the only other simulator capable to handle this) deals with large traces by simplifying the imported map according to its own data format. SUMO provides the tool to import data from online maps (i.e., an OpenStreetMap format), but this operation is highly memory consuming (in our experiments, we were able to cope with scenarios of maximum 100 cars, on a powerful workstation). The problem is that when netconvert transforms an OSM map into a SUMO network file, the size of the generated file becomes considerably larger than the original one. Also, another problem with netconvert is the low level of details which are kept in the resulting SUMO network, with incomplete information about streets, number of lanes on a street, etc. [11].

Currently, almost all tracking devices use GPS technologies, so the data representation in GPS coordinates (WGS84) is a requirement for being a general simulator. SUMO and MATSim use Cartesians coordinates. The direct utilization of mobility traces collected by some GPS receivers in a raw format is impossible, because they firstly need to be processed and converted to the demanding coordinates system of these simulators. When SUMO is connected to the real sensors it manifests an overhead determined by the conversion of coordinates having delays that can affect the traffic flow. MOVE being an extension of SUMO inherits all his problems.

Another problem with analyzed traffic simulators is the fact that their implementations are single threaded, and this radically increases the time of simulation in case of intensive traffic experiments. The authors of [12] confirm this aspect, with an experiment with 10.000 vehicles simulated on the streets infrastructure of Coimbra, Portugal, which lasted 27.000 seconds, an unrealistic amount of time, more than the actual simulated time.

III. ARCHITECTURE

Sim²Car is a simulator designed to provide new research capabilities for next-generation traffic applications. It offers a generic tool for vehicular traffic evaluation: its underlying mobility and networking models are easily extendable; it supports large-scale simulations involving potentially thousands of cars; it offers a realistic mobility model that incorporates real-world data captured in traffic (currently it provides out-of-the-box capabilities for practitioners to evaluate their ITS solutions considering realistic scenarios involving real-world taxis in San Francisco or Beijing); it offers tools to correct uncertainty of location information related to flaws in GPS data sensing and conversion mechanisms to various coordinate systems. The realistic property of its mobility models is boasted by the use of real GIS datasets, freely provided by the OpenStreetMap community. The geographical information has enough details about an urban environment in order to be a good support to generate realistic mobility traces and to simulate complex traffic scenarios. Therefore, developers could construct simulations where cars can sense their environments and take decisions based on the graph of streets declared by the open-source community and already made available within the OpenStreetMap project [13].

Sim²Car includes a tool called TracesTool. It can correct raw data collected by traffic infrastructure sensors (as the data publicly available in real-world data traces – see [14] for a collection of such public data) and make it compatible with the OpenStreetMap digital map format. TracesTool is also capable to generate mobility traces that conform to the real street network provided by OSM, and respect the real traffic flow from GPS datasets.

Sim²Car's architecture is designed to easily integrate and simulate a wide range of traffic scenarios (see Figure 1). Around the simulator core, TracesTool ensures several functions: building street network graph, correction of the raw mobility datasets collects by GPS sensors like [14], [15], [16],

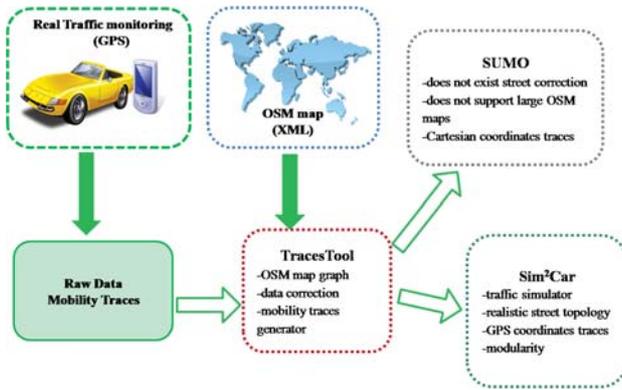


Figure 1. The proposed architecture

generating new mobility traces from corrected traces under the restrictions of OSM graph.

The simulator is designed to support simulation scenarios with overlage street networks. It can generate new synthetic mobility models, starting for example from the original mobility trace. This allows, for example, to modify a simulated vehicular traffic scenario according to different conditions: congestion rate, level of pollution, restricted areas, etc.. Thus, *Sim²Car* is a simulator easily configurable and extendable, for next-generation ITS applications. Figure 1 shows the proposed architecture, together with a possible execution flow. The simulator also offers the possibility to interoperate with SUMO, through resulted mobility traces from TracesTool, modifying them in the SUMO format.

A. TracesTool

TracesTool is a Java-based application that helps users convert the mobility traces into *Sim²Car*'s data format, and deal with uncertainty in GPS sensing – the raw data collected by GPS receivers from cars contains a lot of errors (generally, GPS sensing is known to have a deviation +/-50-100 meters [17] from the real position). TracesTool implements a correction algorithm (described below), to corrects the GPS data. To maintain the realistic characteristic of the original vehicular data, all operations conform to the map graph obtained from OSM map.

The new resulted traces respect the real streets topology, but they do not have an uniform time resolution (due to the periods when GPS receiver does not transmit car position to data collector center). Thus, an improvement of the quality of simulation is also presented below – we use a flow-centric correction algorithm to supply intermediary trace points when necessary.

B. Sim²Car

Sim²Car is next constructed as a discrete time simulator – simulation clock advances with a fixed time unit. The simulator provides capabilities to simulate realistic mobility models, as it combines GIS data from an OSM graph with GPS mobility datasets. The generated traffic closely follows the real-world original mobility conditions.

Figure 2 shows the modular architecture of simulator, and how its components interact. The *Simulation Engine* is the core of the simulator. Its role is to analyze vehicles involved in simulation, and execute their actions for different time moments (the internal clock is advanced with a fixed time resolution).

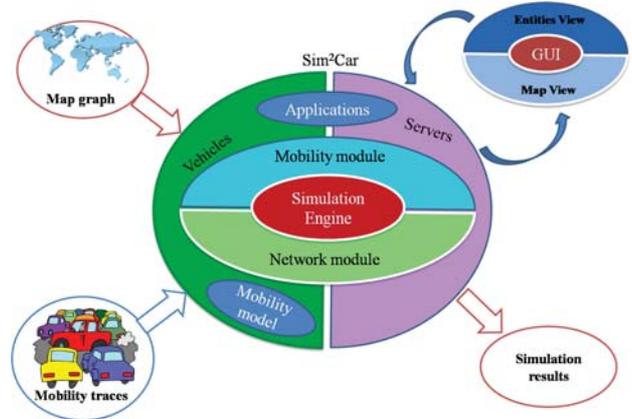


Figure 2. *Sim²Car* Architecture

The *Vehicle* entities, next, model the car behavior. Every car is characterized by a mobility module (in charge with the update of the traffic route), a network module (dealing with communication capabilities, an application layer (implements an application logic, under simulation) and optionally the graphical view.

The *Server* entities model static elements, generally positioned with the street infrastructure, at fixed locations. For example, we can model intelligent traffic lights, Induction Detection Loops, various servers. Each server is characterized by network module (communication capabilities), an application layer (the application logic) and optionally a graphical view.

The *mobility module* updates the position of each vehicle according to the input traces. It follows the traffic rules and positions cars on passing streets. Also, the communication between simulated nodes is sustained by the *network module*, where communication is realized via messages exchange. The network module offers a generic interface which can be particularized to different technologies (i.e., Wireless, Bluetooth, cell-based communication). For example, *Sim²Car* provides a *Wireless network* module, that integrates functions such as: discovering of peers and/or servers, sending and receiving messages (all used by the application layer).

The *application layer* simulates the application running on car's computer (or on the driver's smartphone). Various ITS and traffic applications can be simulated at this layer. Such an application is detailed later on in Section VI.

IV. IMPLEMENTATION

A. TracesTool

TracesTool, written in Java, supports operations for building streets graph, raw datasets corrections and traces generation. It works with several entities: Node, Way, Location.

A point from an OSM map is represented as a *Node* object. It stores details about a POI (Point of Interest), such as the GPS coordinates (latitude and longitude), and the node identifier.

The *Way* object is next used to describe a street parsed from an OpenStreetMap file. It contains the nodes (Node objects) forming the street. We also keep information such as: the neighboring streets which intersects the current one (sorted by junction nodes identifier), street identifier, physical boundaries, and type (one or two directions traffic allowed).

Location describes a point in the mobility trace. The object keeps attributes for a trace point (collected in the real-world trace): geolocation (latitude and longitude), timestamp of collection, information about the vehicle (e.g., for the 500 traces of cabs from San Francisco [14], every record in the trace shows also if the taxi is occupied or not).

We decided to represent the oriented graph of streets as a dictionary of *Way* objects, containing the list of identifiers for all the intersected roads. To increase the searching speed on large maps, we introduced additional levels of indexation, by divided the map surface in squared areas (an edge can be experimentally adjusted according to streets density in the map, and the searching speed increases because the point localization on map becomes a detection of the row and column in the grid).

The algorithm for building the street graph includes two phases:

1) *Parsing Phase* – Initially, the input data of the algorithm is the XML file of a map exported from OpenStreetMap. This raw data is parsed and structured in a format easy to be used in the next step. The resulting file contains all the ways from the map, and for each road we keep the node identifiers and associated tags.

2) *Building Phase* – The previously file is further processed, to detect relations between streets and skip useless OSM elements (i.e., bike paths, pedestrian ways, buildings, and so on). After this step, the algorithm generates several files. A first file contains all vehicular ways. For each street, we store the way identifier, type and all the nodes with associated GPS coordinates. Another file keeps, for each road, the list of junction nodes with correspondent neighbor streets. For this, the algorithm forms a grid on the map, where the edge of a cell is adjusted by the user according to the desired streets density of the map surface. Every cell will further contain the streets which cross through the perimeter delimited by its boundaries.

For the *correction algorithm*, presented below, the Traces-Tool prepares the input using parsed functions adapted to the different GPS raw datasets. The algorithm works correctly with collections of *Location* objects and the *map graph*. To deal with the uncertainty in GPS readings, most GPS correction methods [18] rely on the projection of the GPS reading to the surrounding streets. Our algorithm consider the mobility flow – for each car, each GPS point representing its trajectory are, again, projected on the nearest streets. From the set of candidate corrective points, we select one which conforms to the previous trajectory (i.e., the point that does not disconnect the flow of points on the streets). In this algorithm,

the distance between two GPS coordinates is computing using the haversine formula [13].

```

foreach point in traceData{
/*crtAreaStreets - contains all the streets from the square
  where the point is located on the map
projection(point, street) - return the projection of the
  point on the street
Delta - error range experimentally set between 10 and 100
  meters.*/
crtAreaStreets = getAroundStreets(point);
foreach street in crtAreaStreets{
  dist = dist(point, projection(point, street));
  if( dist < Delta ){
    add(candidates, point, street.id);
  }
}
/* Remove all points with an unacceptable GPS error (no
  candidates)*/
foreach point in traceData{
  if( candidates[point] == [] ){
    remove point from candidates;
  }
}
/* convergence_criteria experimentally established ( the
  remained unresolved nodes ) */
do{
/* Keep the original traffic flow. Correct the points
  which are situated on the wrong way of a street.*/
foreach (crt_point,next_point) in traceData{
  next_point_candidates = candidate[next_point];
  foreach street in next_point_candidates{
    if( angle( direction(crt_point, next_point),
      street_direction ) > 90 ){
      remove candidates[next_point][street.id];
    }
  }
}
/* Removing too far candidates */
foreach point in traceData{
  foreach street in candidates[point] {
    if( dist(point, projection(point, street))> 20) {
      remove street from candidates[point];
    }
  }
}
/*Remove all neighbors that are not situated on the same
  street or two jointed streets.*/
foreach street1 in candidate[crt_point]{
  next_point_candidates = candidate[next_point];
  foreach street2 in next_point_candidates{
    /* junction(street1, street2) tests if exists
      junction between street1 and street2
    if( street1.id != street2.id && !junction(street1,
      street2) ){
      remove candidates[next_point][street2.id];
    }
  }
}
}
Analyze a subset of points from trace data and
if 0.80 of them are on the same street put the
remained points on the same street.
} while( !convergence_criteria );
}

```

Next, we deal with errors caused by the lack of data. In the input mobility trace, it often happens that nodes (cars) failed to send GPS data with a constant rate all the time to the sink server. Thus, our correction algorithm is not always capable to rebuild the original traffic flow all the times (we actually experimented data gaps in the traffic flow in the generated mobility data). To solve this, we next developed another algorithm to deal with generating synthetical mobility points for the missing gaps, such that to have a continuous flow of mobility for each car. This newly generated data trace has to respect in all details the original car itinerary, and the streets network.

Firstly, we determine the minimum global time over all traces. After that, we process the corrected traces in several phases:

Phase 1 – The algorithm first aligns the start time for each trace (data belonging to each car) to a global minimum time.

Phase 2 – The algorithm next analyses the time interval between consecutive points. According to the resulted time gap, the algorithm treats the traffic flow using different methods. These methods try to keep the realistic aspect of the trace conforming to the original vehicular movement. The effort is additionally sustained by the usage of streets topology, since the algorithm fills the gap between points with new entries which are generated according to the trace context: the locations in the roads graph of the analyzed points and the length of time gap.

Method 1 – If the time interval is equal to the fixed resolution, or higher than the superior limit, the current point is kept in the resulted trace. The situation when the time gap exceeds superior limit is met, for example, when the GPS receiver was stopped for a while. In this case, we keep the point in the resulted trace. When the trace is simulated, *Sim²Car* will disable the car drawing on the graphical user interface because of this time gap. The car’s disappearance from the map is called “teleportation”, and is connected to the situation when the GPS does not have signal in places such as tunnels or underground parking.

Method 2 – In this case the value of time difference is between the value of fixed resolution and the superior limit. If the two analyzed points are on the same street, the algorithm adds to the trace new points, computed using their equidistantly place on the line (determined by the two points and their projection on the road).

Method 3 – During this stage, the algorithm treats two points which are on the intersected streets as having similar time issue. Firstly, the junction point is determined. Next, the number of needed points to fill the gap is computed similarly to the method described in *Method 2*. The new points are equally distributed before and after the cross point.

Method 4 – This method deals the situation where the gap between points is too large, so there are several streets between the two consecutive points. For this case, we apply Dijkstra’s algorithm to find the optimal path between them. Before using this algorithm, a subgraph that contains all the streets on a range of a given number of streets is formed around the first point. If the algorithm finds a path between the two points, the algorithm next places new points on every street of the path, with an uniform spatial step. For the Dijkstra’s algorithm we use as weight the length of the street.

The traces generator offers a generic format for the resulted traces. One record from the output trace contains the following information about the vehicle at a certain moment: latitude, longitude, timestamp, street identifier, and other details. Adjusting with a format convert, a trace generated in the above way can be used in SUMO having the advantage that the trace points will be placed on the right streets and the disadvantage that SUMO will probably not be able to load the graph of a

large map like the one of San Francisco city in the RAM of a personal computer.

B. *Sim²Car* Implementation

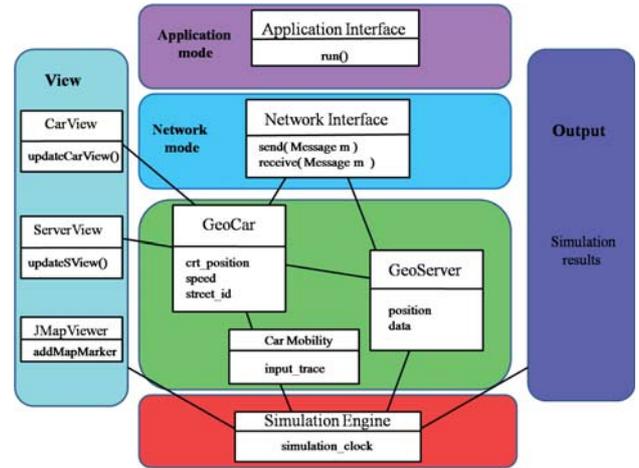


Figure 3. UML diagram of *Sim²Car*.

Sim²Car is developed as a modular simulator. Figure 3 presents an UML diagram of *Sim²Car*, showing the main components. At the core is the *simulation engine*. On top, the behavior of the nodes (vehicles, servers) is modeled by the *Entity* class. Depending on the simulated scenario, different types of entities can exist, such as cars and servers.

The mobility is management by the *CarMobility* class, in charge with advancing the current position for each car entity. As an optimization, the class does not load an entire car’s trace into memory. Instead, it only reads the car’s current position and the next position from the trace file.

Next, the *network module* is a member of the *Entity* class, because any car and server can have one or more network interfaces. This module is easily extensible, as it can be adjusted to different technologies: 3G/4G/WiMAX, Wireless or Bluetooth, IEEE802.11p, etc. In the current simulator we implemented a Wi-Fi module, which offers through its API methods for discovering simulation nodes and provides communication primitives (using message exchanges). A message is modeled by the *Message* class. In general, the network layer is used at the application layer to communicate to other nodes in proximity, having the same application installed (running in the simulation).

The *Application* interface models the logic at the application layer. By implementing this interface, anyone can add inside the simulation new ITS applications and services: to monitor traffic flow, congestion, routing, etc.. In Section VI we present an example of such an application.

On top, the Graphical User Interface is implemented through a *View* class. The class operates three main areas: map area, control area and logging area. For the slippery map, for example, we use the *JMapView* [19] library, provided by OpenStreetMap community. *JMapView* is a useful tool for rendering an OSM map using online resources or downloaded

files. Being an open-source project, it was easily adjustable to our needs, and comes already equipped with a vast collection of GUI features. For example, the GUI currently provides integrated zoom controls, the possibility to add markers and polygons on the represented surface. The control area can be used by the users to control the simulation flow. The logging area shows different information about actions performed by the simulated nodes.

V. EXPERIMENTS

A. OSM Graph Building Algorithm

For evaluating our corrective algorithms for the mobility data traces, we choose five maps with different sizes from Metro Extracts site [20]. The site provides parts of the OpenStreetMap database for the majority of world cities. We first converted the OSM maps into SUMO network format using netconvert. Next, we converted the same maps into OSM graph format, using our algorithm. A comparison between the resulting graphs and indexation table (in terms of file size, and memory needs to load the data), is presented in Tables I (for SUMO) and II (for our algorithm). The measurement unit is MB, and the acronyms are: *XFS* - XML File Size, *LSNFS* - Large SUMO Network File Size, *RSNFS* - Reduced SUMO Network File Size, *AFS* - Areas File Size, *SFS* -Streets File Size, and *AdFS* - Adjacency File Size.

During our tests, we noticed several debatable situations such as: warnings, unknown OSM XML tags, problems with physical memory. Besides, netconvert demands to edit the original OSM file with the JOSM application (otherwise it does not make differences between pedestrian ways and vehicular street, lanes), and this requires an extra effort for the users.

The map editing on personal computers is not possible all the times for complex maps (e.g., *San Francisco*, involving 500 taxis), because the JOSM tool is not able to load the entire map into memory, or if it succeeds, the experience is quite annoying for the users due to the lack of application interactivity.

After the editing phase, all maps were accepted by netconvert as input, but results were unsatisfactory because the output file size is almost double than original file size. The OSM maps use GPS coordinates for internal elements, while SUMO uses Cartesians coordinates system for data representation (keeping additional metadata for in its internal orientated graph). Therefore, the conversion between both coordinates systems can increase the resulted file size.

We used the same tests for our OSM graph building algorithm. The *Areas File* represents the file which keeps the indexation table for the entire map. The *Streets File* contains all streets from OSM map with their nodes. The *Adjacency File* contains for each street from OSM map lists of neighbors and every list is identified by junction point identifier.

During the experiments with our corrective algorithm, we did not encounter problems. Even for the *Paris* map, the algorithm managed well to handle the load into memory. As the results in Table II show, we manage to utilize an

Table I
THE RESULTS OF SUMO TESTS

Map	XFS	LSNFS	RSNFS	RAM Used
San Francisco	255	629	500	≥ 1930
Shanghai	195	509	383	≥ 1723
Paris	3660	OofM	OofM	-
Bucharest	0.876	1.75	1.32	21
Brussels	544	801	456	≥1954

Table II
THE TESTS RESULTS FOR OSM GRAPH BUILDING ALGORITHM

Map	XFS	AFS	SFS	AdFS	RAM Used
San Francisco	255	0,982	21,4	9,28	137
Shanghai	195	5,879	27,428	11,019	178
Paris	3660	3,08	58,3	30,03	240
Bucharest	0.876	0,003	0,055	0,033	8
Brussels	544	1.31	24	12.6	175

acceptable quantity of the memory, and still offer to the user the possibility to use streets graph in *Sim²Car* without leading to memory problems.

VI. USE CASE SCENARIO: TILES APPLICATION

To evaluate *Sim²Car*'s capabilities, we develop on top an application (called *Tile Application*) designed to optimize the storing of location-aware data in a medium composed of fixed points (workstations), wireless access points and smartphone applications (as in typical smart city scenarios). This application, demonstrated in [21], uses context information (location, neighbors) and techniques specific for opportunistic computing to provide an efficient management of the available data for satellite navigation systems based on high resolution raster maps.

We simulated different scenarios using mobility datasets collected in San Francisco [14] and Beijing [15], [16]. The territory of San Francisco, California covers an area of about $121 km^2$ [22], whilst Beijing encompasses an area of about $16.807,8 km^2$ [23]. For the *Tile Application* we used the following parameters: each Client (car's computer, or smartphone) has a local storage capable to hold up to 5 MB (average space occupied by modern apps). This data is represented by tiles (geographical data needed for the representation of the navigation system). Each tile covers a geographic area of $0,4 km^2$, for San Francisco, and $6 km^2$, for Beijing. The user can download tiles from neighbors (if available) directly over wireless close-range communication (which actually saves transfer costs). Otherwise, if a tile is not available, the user has to download a tile from a central repository. Each tile has a resolution of 256256 pixels. We actually used real map tiles downloaded from the OSM site, and considered that each contact between simulated entities (car-to-car, car-server) last sufficient to download the necessary tiles. For the communication between cars, we simulated the use of 802.11p technology (with a range of approx. 300 meters, and maximum data throughput somewhere between 6 and 9 Mbit/s) [24].

We present below the results for two scenarios. First, we simulated the movement of all 500 cabs in the San Francisco datasets, over the period of 12 days (from 17th May 2008 to 29th May 2008). In this case *Sim²Car* models both cars and servers (the *Tile Application* logic was implemented also for the central repository of tiles, aka the server). We defined 10 servers (wireless routers capable to provide tiles), uniformly spread over the area of San Francisco (we used the main junctions actually). The entire simulation lasted 19.020 seconds (approximately 5 hours and 30 minutes).

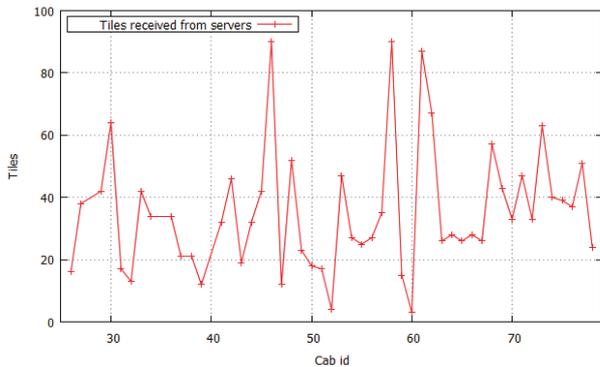


Figure 4. The number of tiles receive by peers from servers (San Francisco).

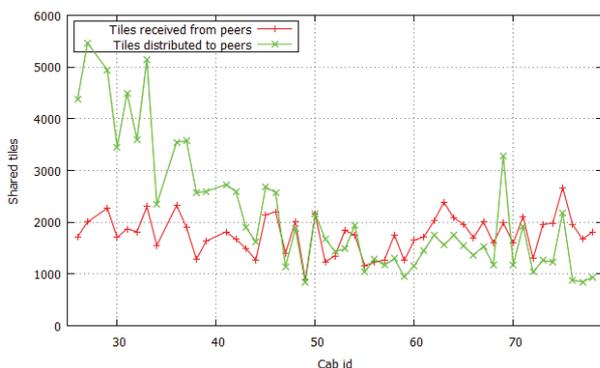


Figure 5. The number of tiles exchange between peers (San Francisco).

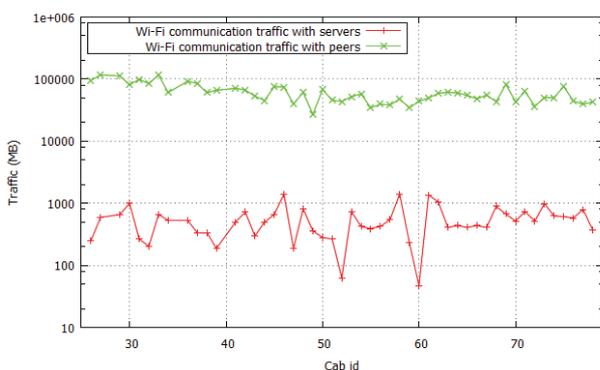


Figure 6. The network traffic over Wi-Fi (San Francisco).

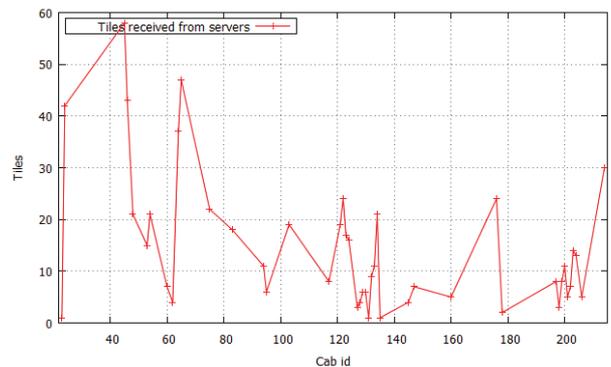


Figure 7. The number of tiles exchange between peers (Beijing).

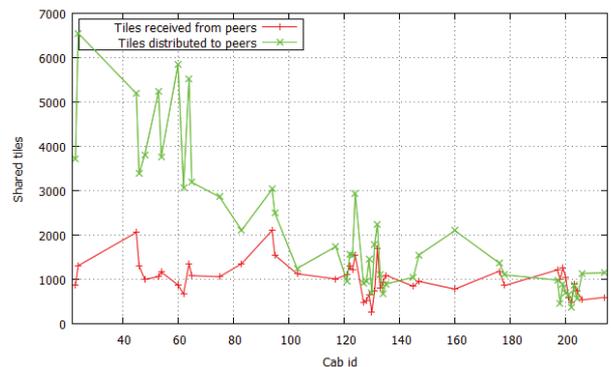


Figure 8. The number of tiles exchange between peers (Beijing).

We randomly selected 50 cabs from all simulated cabs, and plotted the data collected in the simulation. Figure 4 shows the number of the tiles received by peers from servers, and Figure 5 presents the number of tiles provided and the ones obtained to/from peers. The results show that the application can actually lead to a reduction in traffic, which is especially valuable in crowded urban environments. In this case, data can be better disseminated between peers, reducing the expenses of the drivers and the network traffic to central servers. As the results show, peers seldomly established connections to servers to require tiles.

Figure 6 shows the Wi-Fi traffic realized in peer-to-peer and client-to-server communication. In this example, peers send to other peer/server the ID of the required tile (approx. 8 Bytes). A sent/received tile has the cost of 20 KBytes. As seen, peers generally take the needed tiles from the other peers – which is the main reason why peer-to-peer network traffic over Wi-Fi reaches a GB order of magnitude.

In the second set of experiments, we simulated the movement of 1000 cabs from Beijing datasets, during one week (from 2nd February 2008, to 10th February 2008). We added 55 servers, again uniformly spread over the area of Beijing (in its main junctions). The simulation lasted 38.200 seconds (approximately 10 hours and 30 minutes). We also used the same costs for network traffic over Wi-Fi as in the previous set of experiments.

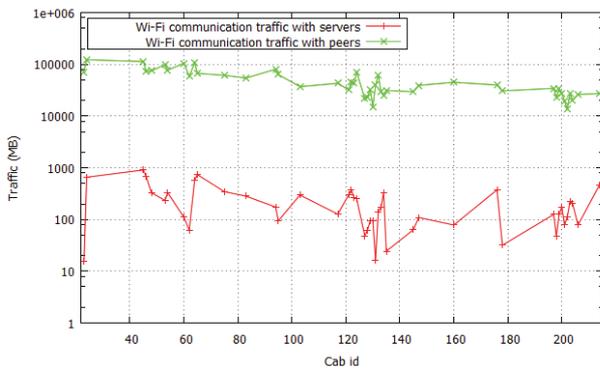


Figure 9. The network traffic over Wi-Fi (Beijing).

Figures 7, 8, and 9) show the obtained results. The *Tiles Application* keeps again the traffic over Wi-Fi in similar proportions as in the previous experiments. However, in this case we witness a decrease in the client-to-server traffic over Wi-Fi, which is due to the high density of peers involved in traffic.

VII. CONCLUSION

In this paper, we presented *Sim²Car*, a generic tool for vehicular traffic evaluation which supports large-scale simulations involving potentially thousands of cars. Currently, our simulator provides a realistic mobility model that incorporates real-world traces (San Francisco and Beijing datasets). For the input data, we constructed our tools to correct raw GPS datasets according to streets graph, using a mapping algorithm.

We evaluated the simulator's capabilities under an application, *Tiles Application* that we previously demonstrated in [21]. Under extensive experiments in *Sim²Car*, we demonstrate the capability of the simulator to handle a large number of nodes – we successfully simulated large traffic scenarios, involving 500 cabs in San Francisco, and 1000 cabs in Beijing. *Sim²Car* manages very well such kind of experiments, being capable to handle thousands of cars and large OSM maps, covering hundreds of km^2 .

ACKNOWLEDGMENT

The research is supported by CyberWater grant of the Romanian National Authority for Scientific Research, CNDF-UEFISCDI, project number 47/2012. The work was also supported by the project "SideSTEP - Scheduling Methods for Dynamic Distributed Systems: a self-* approach", PN-II-CT-RO-FR-2012-1-0084.

REFERENCES

- [1] C. Systematics, *Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation*. Federal Highway Administration, 2005, vol. 6.
- [2] C. Fratila, C. Dobre, F. Pop, and V. Cristea, "A transportation control system for urban environments," in *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on*. IEEE, 2012, pp. 117–124.
- [3] A. Gainaru, C. Dobre, and V. Cristea, "A realistic mobility model based on social networks for the simulation of vanets," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*. IEEE, 2009, pp. 1–5.

- [4] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," 2013.
- [5] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3 and 4, pp. 128–138, 2012.
- [6] (2014, Feb.) Simulation of urban mobility. [Online]. Available: <http://sumo.sourceforge.net>
- [7] (2014, Feb.) Move project homepage. [Online]. Available: <http://lens.csie.ncku.edu.tw>
- [8] A. K. Banik, "Routing protocol with prediction based mobility model in vehicular ad hoc network (vanet)," 2010.
- [9] V. Cristea, V. Gradinescu, C. Gorgorin, R. Diaconescu, and L. Iftode, "Simulation of vanet applications," *Automotive Informatics and Communicative Systems*, 2009.
- [10] (2014, Feb.) Vnsim homepage. [Online]. Available: <http://cipsm.hpc.pub.ro/vanet/vnsim.html>
- [11] A. K. Mario Krumnow, "Simulation of vanet applications," *Real-time simulations based on live detector data Experiences of using SUMO in a Traffic Management System*, 2013.
- [12] D. C. S. Jos Capela Dias, Pedro Henriques Abreu, "Simulation of vanet applications," *Preparing Data for Urban Traffic Simulation using SUMO*, 2013.
- [13] C. Vetter, "Fast and exact mobile navigation with openstreetmap data, diploma thesis," 2010.
- [14] (2014, Feb.) Crawdad community, san francisco bay traces download section. [Online]. Available: <http://crawdad.cs.dartmouth.edu/meta.php?name=epfl/mobility>
- [15] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. ACM, 2010, pp. 99–108.
- [16] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 316–324.
- [17] C. D. Dragoi Vlad Alexandru, "Smart traffic, master thesis," 2012.
- [18] J. E. Robbins, "Gps correction methods, apparatus and signals," Sep. 28 2004, uS Patent 6,799,116.
- [19] (2014, Feb.) Jmapviewer project homepage. [Online]. Available: <http://wiki.openstreetmap.org/wiki/JMapViewer>
- [20] (2014, Feb.) Metro extracts home. [Online]. Available: <http://metro.teczno.com>
- [21] D. Urda, C. Dobre, and F. Pop, "Storing location-aware data in mobile distributed systems," in *Parallel and Distributed Computing (ISPD), 2013 IEEE 12th International Symposium on*. IEEE, 2013, pp. 135–142.
- [22] (2014, Feb.) Us census bureau, state and county quickfacts, san francisco (city), california. [Online]. Available: <http://quickfacts.census.gov/qfd/states/06/0667000.html>
- [23] (2014, Feb.) Beijing: Population and density by district and county. [Online]. Available: <http://www.demographia.com/db-beijing-ward.htm>
- [24] (2014) Mit technology review, the internet of cars is approaching a crossroads. [Online]. Available: <http://www.technologyreview.com/news/515966/the-internet-of-cars-is-approaching-a-crossroads>

**Research and Use of
Multiformalism
Modeling Methods
-
Special Session**

MULTI-FORMALISM MODELING FOR EVALUATING THE EFFECT OF CYBER EXPLOITS

Alexander H. Levis and Bahram Yousefi
System Architectures Laboratory
George Mason University
Fairfax, VA USA

E-mail: alevis@gmu.edu; byousefi@masonlive.gmu.edu

KEYWORDS: Decision making organizations, cyber exploits, synchronization, information sharing

ABSTRACT

An approach based on multi-formalism modeling to model, analyze, and evaluate the effect of cyber exploits on the coordination in decision making organizations is presented. The focus is on the effect that cyber exploits have on information sharing and task synchronization. The organization members are supported by systems and interact with each other through communication networks. Colored Petri Nets are used to model the decision makers in the organization and computer network models to represent their interactions. Protocols of interaction are modeled by rules of enablement where the decision makers, when interacting, must refer to the same state of the environment. Two measures of performance are then introduced: information consistency and synchronization. The multi-formalism based modeling approach and the computation of the measures of performance are illustrated through a simple example.

INTRODUCTION

Assessing the effect of cyber exploits on an organization's performance is a challenging problem. A cyber exploit is an action that affects the performance of an information system by taking advantage of its cyber vulnerabilities. The evaluation of the effectiveness of a decision making organization consisting of human decision makers supported by systems and interacting through networks is a complex issue: many interrelated factors affect the effectiveness of the overall system, e.g., the limited information processing capacities of the decision makers and the hardware and software characteristics of the systems. Consequently, models are needed of organizations performing well defined tasks and of their information systems, as well as performance evaluation measures and procedures for computing them. An integrated methodology that exploits multi-formalism modeling and is based on some earlier work has been developed and is described in this paper.

One of the key effects of cyber exploits is the degradation of the cohesiveness of organizations carrying out well-defined tasks in a coordinated manner. A

mathematical description of coordination was developed for decision-making processes by Grevet and Levis (1988). When confronted with a particular task, organization members need to access information from the supporting systems and to interact with each other following well defined processes. Such is the case in command centers such as Air Operations Centers, Air Traffic Control centers, etc. When decision makers interact, they must have some protocol to recognize that they are working on the same task and that they are sharing information that pertains to that task. Two measures for evaluating coordination were introduced: *information consistency* and *synchronization*. The latter measure relates to the value of information when the decision makers actually process it.

The approach taken in this work is that of modular, horizontal multi-formalism (Gribaudo and Iacono, 2014). A generic Petri Net model of an interacting decision maker is used. (Levis, 1992) That model has been extended to include systems that support the decision makers and communication networks that enable their interaction. The decision making organization is modeled as a Colored Petri Net (Jensen and Kristensen, 2009) and is implemented in CPNTools (CPNTools, 2014). The computer networks are modeled as queuing nets and implemented in OMNeT++ (OMNeT++, 2014). These two models, each expressed in a different modeling language (formalism), interoperate through an infrastructure, the Command and Control Wind Tunnel. This is shown in Fig. 1. (Hemingway et al., 2001)

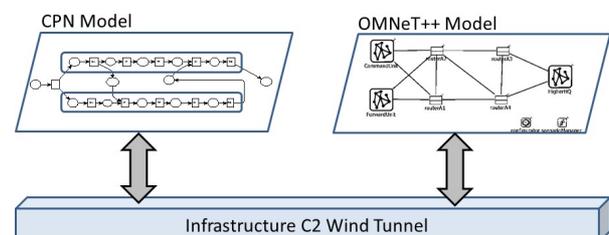


Figure 1: Multi-formalism Architecture

The validity of the interoperation of these two formalisms was established in Abu Jbara and Levis (2013) based on the approach described in Levis et al. (2012).

THE DECISION MAKING ORGANIZATION MODEL

The decision making organizations under consideration consist of groups of interacting decision makers processing information received through systems that enable information sharing (e.g., a cloud) and who interact to produce a unique organizational response for each task that is processed. Each interacting organization member is modeled as consisting of a five-stage process as shown in Fig. 2.

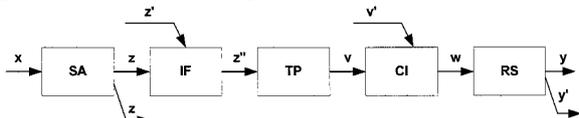


Figure 2: The Five Stage Decision Maker Model

The decision maker receives a signal, x , from the external environment or from another decision maker. The Situation Assessment stage (SA) represents the processing of the incoming signal x to obtain the assessed situation, z , which may be shared with other decision makers. The decision maker can also receive situation assessment signals z' from other decision makers within the organization; z' and z are then fused together in the Information Fusion (IF) stage to produce z'' . The fused information is then processed at the Task Processing (TP) stage to produce v , a signal that contains the task information necessary to select a response. Command information from other decision makers is received as v' . The Command Interpretation (CI) stage then combines v and v' to produce the variable w , which is input to the Response Selection (RS) stage. The RS stage then produces the output y to the environment, or the output y' to other decision makers.

A Petri Net model is used to depict interactions between decision makers; the admissible interactions are limited to the four types shown in Fig. 3 in which only the interactions from the i^{th} (DM_i) to the j^{th} decision maker (DM_j) are shown. Similar interactions exist from the j^{th} to the i^{th} one. Furthermore, not all these interactions can coexist, if deadlocks are to be avoided (Remy and Levis, 1988).

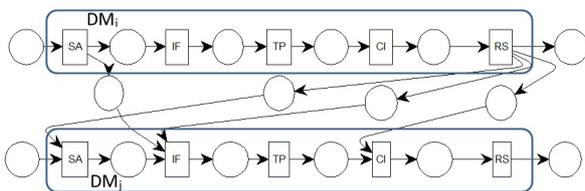


Figure 3: Interacting Decision Making Entities

Figure 4 shows an example of a decision making organization consisting of a forward unit and a command unit that can perform three different missions: humanitarian assistance, disaster relief, or a non-combatant evacuation. Both units assess the signals that they receive from the environment in their respective Situation Assessment stages. The forward unit (subordinate) sends the

result of its own assessment to the command unit (commander), who fuses in the Information Fusion stage this information with its own assessment. On the basis of the result of this interaction, the command unit identifies the mission or situation and produces an order which is sent to the subordinate unit. The latter interprets the order in the Command Interpretation stage and executes the mission.

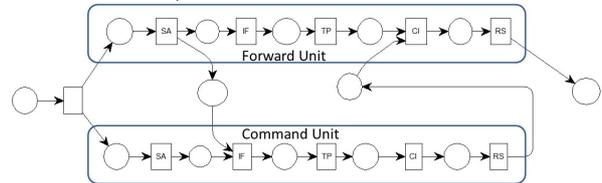


Figure 4: A Two Unit Organization

A decision maker may have access to or select different systems and different algorithms that process the input depending on the type of signals received. The DM chooses an algorithm according to his area of expertise and the prevailing circumstances (timeliness, access to systems, etc.). Each algorithm is characterized by the accuracy of its output and the associated delay in producing it. The algorithms may all reside in one system (e.g., at the command unit) or be located in different system nodes; they may be accessible directly through an intranet or they may be accessible through the communication networks. There may be inconsistent or conflicting assessments at the IF stages of the two units for a variety of reasons: using different data sets (e.g., new vs. old data) or different assessment algorithms. A mechanism would be needed to resolve such inconsistencies or conflicts. A similar argument is made about the Response Selection stage where DMs have different algorithms for generating a response. The expanded model of the example (Fig. 5) shows the existence of three SA algorithms and three RS algorithms to reflect the three different missions that the organization has been trained to execute. This is a Colored Petri Net model (Jensen and Kristensen, 2009) with deterministic time but stochastic algorithm selection. The application CPN Tools (2014) has been used to implement the model.

The organization in Fig. 5 is a team of two DMs that collaborate with each other to make decisions based on the available inputs.

1. They share the same goal(s), i.e., the set of output responses is shared between all the team members.
2. They share the same skill(s) where the skill set is modeled by the sets of three algorithms in the SA stage and the three in the RS stage. This means that they have the same sets of algorithms but, because they have different areas of expertise, they may use their skill sets in different ways.
3. They should be synchronized to be able to perform tasks effectively.

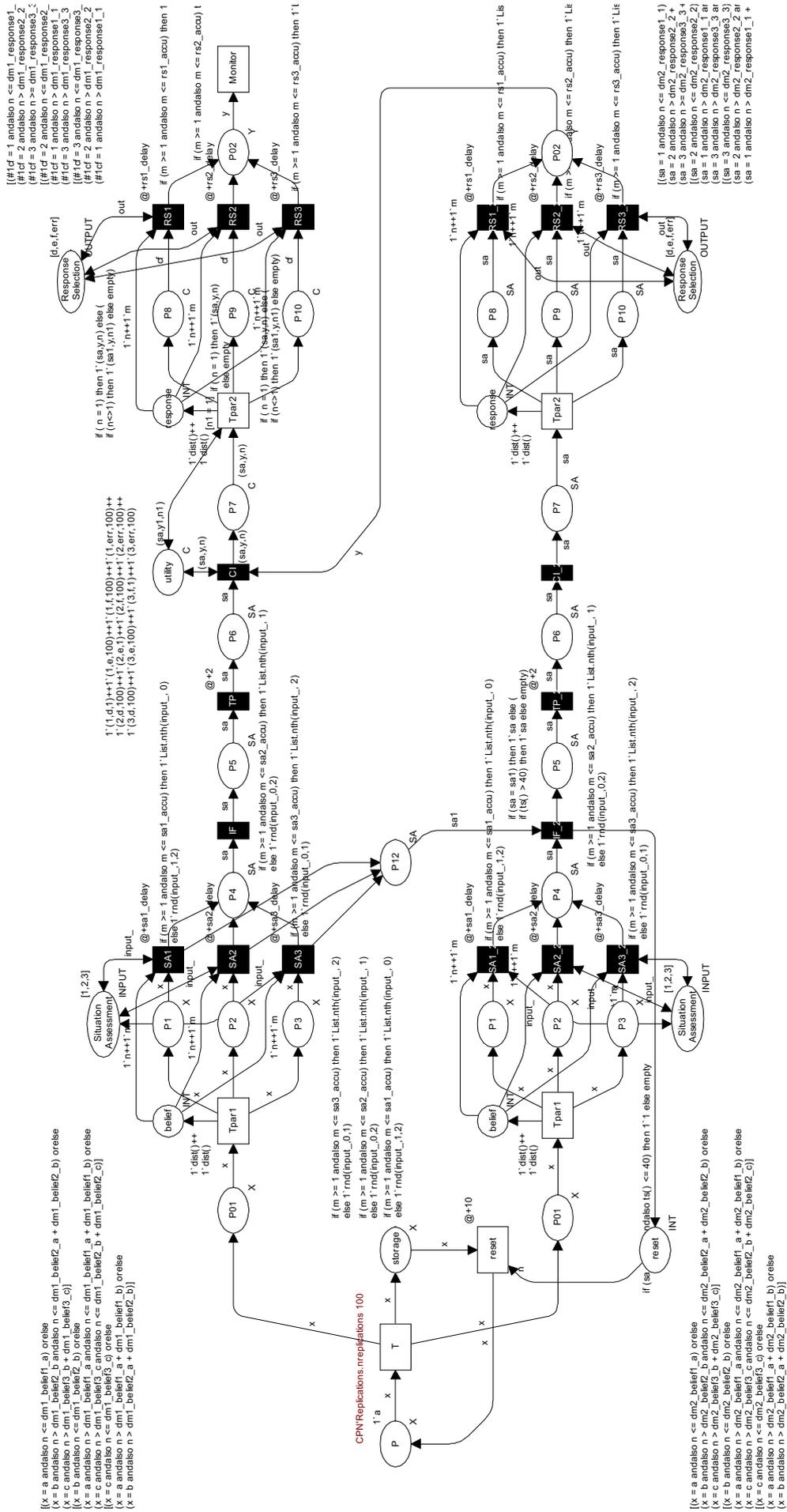


Figure 5: The Detailed Organization Model

ON MEASURES

In order to assess the effect of cyber exploits, measures are needed. A set of new measures is defined here that is computable from the Colored Petri net model of the decision making organization. These measures were originally defined in Grevet and Levis (1988).

To characterize the coordination for an interaction such as at the IF or CI stage in the model of Fig. 5 order relations are defined on the set of tokens fired by the corresponding transition:

Ψ_1 is a binary relation defined by:

$$((x, y, z) \Psi_1 (x', y', z')) \leftrightarrow ((x = x') \text{ and } (y \leq y'))$$

Ψ_2 is a binary relation defined by:

$$((x, y, z) \Psi_2 (x', y', z')) \leftrightarrow ((x = x') \text{ and } (z = z'))$$

Each token in the Colored Petri Net model is characterized by the triplet (T_n, T_d, C) where T_n is the time at which this input token was generated by the source, T_d is the time at which a token entered the current processing stage, and the attribute C characterizes the mission or task.

The firing of IF (or CI) is synchronized if and only if:

$$\forall i \in \{1, 2, \dots, r\} \quad (T_n^i, T_d^i, C^i) \Psi_1 (T_n^k, T_d^k, C^k)$$

where i and k are two decision makers. This definition allows to discriminate between firings that are synchronized and firings in which one or several tokens arrive in their respective corresponding places with some delay.

The firing of IF (or CI) is consistent if and only if:

$$\forall (i, j) \in \{1, 2, \dots, r\} \times \{1, 2, \dots, r\}, \quad (T_n^i, T_d^i, C^i) \Psi_2 (T_n^j, T_d^j, C^j)$$

i.e., the data fused by a decision maker are consistent if they correspond to the same task or mission C . On this basis, the following definition for the coordination of an interaction is obtained: *The firing of a transition (such as IF) is coordinated if, and only if, it is synchronized and consistent.*

The definition of coordination applies to a single interaction. The definitions of the coordination of a single task, i.e., for a sequence of interactions concerning the same input, as well as for all tasks executed in a mission are as follows: The execution of a task is coordinated if, and only if, it is coordinated for all interactions that occur during the task. The execution of a mission is coordinated if, and only if, it is coordinated for all its tasks.

Consider a transition such as IF (or CI) with multiple input places. The $V(x_i, IF)$ denotes the vector that describes the colors of the tokens in the preset that have been generated as a result of the signal x_i (task or mission) produced by the external source. Then the degree of information consistency (DIC) for stage IF and input task x_i is defined as:

$$d(x_i, IF) = \sum_{V(x_i, IF)} \text{prob}(V(x_i, IF)) \frac{n(V(x_i, IF))}{z(V(x_i, IF))}$$

where $\text{prob}(V(x_i, IF))$ is the probability of having tokens with attributes generated by x_i in the input places of IF. Then let z be the number of subsets of two elements of $V(x_i, IF)$:

$$z(V(x_i, IF)) = \binom{r}{2} = \frac{r!}{2!(r-2)!}$$

and let n be the number of subsets of $V(x_i, IF)$ such that the two elements are equal.

By adding the degrees of information consistency for IF and CI and each task x_i and weighing by the probability of having that input task, the organizational degree of information consistency, DIC, for the tasks at hand can be evaluated:

$$DIC = \sum_{x_i} \text{prob}(x_i) \sum_{B=IF, CI} d(x_i, B)$$

This measure varies between 0 and 1, with 1 being the ideal information consistency of all interactions across all tasks.

The total processing time for a task by a decision maker consists of two parts: (a) the total time during which the decision maker actually carries out the task; and (b) the total time spent by the information prior to being processed. The latter time is due to two factors: (i) Information can remain unprocessed until the decision maker decides to process it with a relevant algorithm. Since an algorithm cannot process two inputs at the same time, some inputs will have to remain in queue for a certain amount of time until the relevant algorithm is available. (ii) Information can also remain unprocessed because the decision maker has to wait to receive data from another organization member. Consequently, an organization is not well synchronized when decision makers have to wait before receiving the information that they need in order to continue their task processing. Conversely, the organization is well synchronized when these lags are small.

The degree of synchronization for the organization, DOS, is given by:

$$DOS = \sum_{x_i} \text{prob}(x_i) \sum_{B=IF, CI} S(x_i; B)$$

where $S(x_i, B)$ is the total delay in transition B because of differences in the arrival time of the enabling tokens in its preset.

Two more measures were defined for evaluating the effect of cyber exploits on organizational performance.

Accuracy of the Organization $J(\delta)$ is the degree to which the organization produces desirable results (with

lowest penalty) when using strategy δ . This is a global measure which ideally would be one but in realistic situations it is always less than one.

$$J(\delta) = \sum_i prob(x_i) \sum_h cost(y_h, y_{di}) prob(y_h | x_i)$$

Where $\{x_i\}$ is the set of tasks and $\{y_h\}$ is the set of admissible responses from which the response y_h is selected for task x_i and y_{di} is the ideal response for input x_i . $Cost(y_h, y_{di})$ represents the cost associated with the organization's response.

Timeliness of the Organization $T(\delta)$ is the total response time of the organization from the time a task arrives to the time a response is produced, i.e., the task has been executed using strategy δ .

$$T(\delta) = E(\text{elapsed time})$$

Up to this point, the model of an organization executing a set of tasks that arrive according to some probability distribution has been described. Also, measures of performance of the organization have been defined.

When fusion of data is performed by a decision maker it is possible that the available markings may allow multiple enablement of the IF (or CI) transition. Consequently, enablement rules need to be introduced at this point. Two alternative rules have been considered:

Rule 1: Transition IF (or CI) is enabled, if all its input places contain a token with the same value of the time attribute T_n . Rule 1 means that the transition IF (or CI) is enabled if and only if all its preset places contain at least a representation of the same input x_i .

Rule 2: The transition IF (or CI) is enabled if Rule 1 applies or if delays in receiving inputs from other organization members exceed a pre-specified limit.

THE NETWORK MODEL

In the example of Fig. 5, interactions between the forward unit and the command unit occur through computer communications networks. The network model shown in Fig. 6 used in this example is a free-style wired, packet switching network with TCP/ IP as its protocol suite. This connection-oriented protocol was employed in order to handle reliability, message ordering and streaming at the Transport Layer rather than the Application Layer. The Application Layer will extract the transmitted data and deliver it to the organizational model through the C2 Wind Tunnel (Hemingway et al., 2011) at the designated terminal. The C2 Wind Tunnel (Fig. 1) is a test bed that enables the interoperability of models by providing the necessary infrastructure as well as the tools for scheduling the interactions among inter-operating models. The data received at the terminal is then examined for timeliness and integrity back in the CPN organizational model.

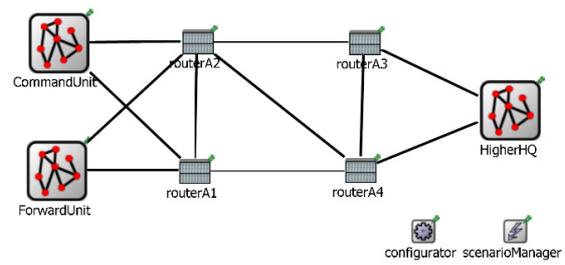


Figure 6: OMNET++ Representation of the Communication Network

The network model was implemented using OMNeT++ (2014) which is an extensible, modular, component-based C++ simulation library and framework, primarily for building network simulators. This is a discrete event simulation environment consisting of simple/compound modules with each module having a defined functionality (according to the relative C++ class). Each module could be triggered with an appropriate event defined in its class. For the model of Fig. 6 the INET (2014) framework was used which supports different networking protocols including the four layers of the TCP/ IP protocol. The framework provides simple/ compound modules for all the four layers of the TCP/ IP. The nodes in Fig. 6 contain in them the internal network structure for the corresponding entity (command unit, forward unit, and Higher HQ).

MODELING CYBER EXPLOITS

Two types of cyber exploits were implemented for the computational experiments: 1) Denial of Service attack and 2) Integrity attack. The Denial of Service is an attempt to make a network resource unavailable or render it too slow to be useful. This attack affects a localized region of the network topology, e.g., some routers. The Integrity attack, as was defined in this case, involves tampering with the contents of the data packets in order to compromise the performance of the organization through deception. It is usually the case that the attacker intercepts the data being passed between two terminals for a considerable amount of time; this sometimes leads to the detection of an anomaly

The two types of attack were implemented in OMNeT++ for a predefined scenario that can generate coordinated attacks of both types. The Denial of Service exploits were modeled as a delay in a communication path or as a total failure of a network resource (e.g., a router) for a finite amount of time as defined in the scenario. The integrity attacks were implemented as an alteration in the message contents, i.e., by changing the attribute values of the tokens, at specified times as defined in the scenario. The results of the attack were evaluated using the measures of performance defined in the measures section.

COMPUTATIONAL EXPERIMENT AND RESULTS

The organization model (Fig. 5) implemented in CPN Tools, and the communication network model (Fig. 6) implemented in OMNet++ inter-operated through the infrastructure provided by the C2 Wind Tunnel (Fig. 1). The scenario generated tasks $\{x_i\}$ to be performed by the organization. As described earlier, three different types of events were considered: Humanitarian Assistance (Event 1), Disaster Relief (Event 2) and Non-combatant evaluation operations (NEO) (Event 3). The area of expertise of each organization member was modeled as the probability of choosing the right algorithm in the Situation Assessment stage that matched the event, and the probability of choosing the right algorithm in the Response Selection stage for the corresponding assessed situation. The values for the Forward Unit are shown in Tables 1 and 2 and for the Command Unit in Tables 3 and 4. For example, the Forward Unit has a 90% probability of assessing correctly a Humanitarian Assistance task but only 80% a Disaster Relief one.

Table 1. Probability of Forward Unit (DM) selecting each SA algorithm for each incoming event

	SA Algorithm 1	SA Algorithm 2	SA Algorithm 3
Event 1	0.9	0.05	0.05
Event 2	0.1	0.8	0.1
Event 3	0.2	0.1	0.7

Table 2. Probability of Forward Unit (DM) selecting each RS algorithm for each situation

	RS Algorithm 1	RS Algorithm 2	RS Algorithm 3
Situation 1	0.9	0.05	0.05
Situation 2	0.1	0.8	0.1
Situation 3	0.2	0.1	0.7

Table 3. Probability of Command Unit (DM) selecting each SA algorithm for each incoming event

	SA Algorithm 1	SA Algorithm 2	SA Algorithm 3
Event 1	0.95	0.04	0.01
Event 2	0.05	0.9	0.05
Event 3	0.05	0.1	0.85

Table 4. Probability of Command Unit (DM) selecting each RS algorithm for each situation

	RS Algorithm 1	RS Algorithm 2	RS Algorithm 3
Situation 1	0.95	0.04	0.01
Situation 2	0.05	0.9	0.05
Situation 3	0.05	0.1	0.85

The cost function included in the definition of the Accuracy measure is shown in Table 5. Selecting the correct response has a cost of 1 while selecting the other two responses has a cost of 100.

Table 5. Cost of Selecting a Response for each Situation

	Response 1	Response 2	Response 3
Situation 1	1	100	100
Situation 2	100	1	100
Situation 3	100	100	1

There is a timer associated with each IF transition. When the first token is placed in one of the input places of the IF Transition, the timer starts. In this simulation the timer's timeout is 40 units of time. Three alternative operational rules were considered:

Rule 1: Transition IF is enabled if all its input places contain a token.

Rule 2: Transition IF is enabled if Rule 1 is not satisfied and the timer times out.

Rule 3: Transition IF is enabled if Rule 1 or Rule 2 applies.

In the Task Processing (TP) stage, the required tasks are performed prior to generating a command or a response. This stage is implemented as a processing delay. The computational experiments were carried out in two different modes:

Single-mode: In this mode, the model was run once for the given scenario. Each run may include one attack either on the IF or CI stage.

Batch-mode: In this mode the model was run in batches of 100 runs with probability 0.2 of an attack (either on IF or CI stage) happening in a single execution

The results of the computational experiments are shown in Tables 6 to 9. Table 6 establishes the baseline for this scenario for the four measures, Accuracy, Timeliness, Degree of Information Consistency (DIC), and Degree of Synchronization (DOS). In this case, there are no delays beyond the minimum processing delays and there are no errors in selecting the appropriate algorithm. The batch mode results show the degradation in performance when there is 20% probability that an attack will take place. Tables 8 and 9 show clearly that integrity attacks at the Information Fusion stage have significant impact on the timely response of the organization and, as a result, the synchronization degrades significantly. An attack on the Command Interpretation stage has a considerable impact on the organization's accuracy in the response to the task.

Table 6: Single-Mode Results: Ideal Organization

Accuracy	Timeliness	DOC	DOS
1.0	8	1	2

Table 7: Batch-mode results

Accuracy	Timeliness	DOC	DOS
0.8	11.28	0.89	4

Table 8: Single-mode integrity attack at IF stage

Accuracy	Timeliness	DOC	DOS
1.0	44	0.625	40

Table 9: Single-mode integrity attack at CI stage

Accuracy	Timeliness	DIC	DOS
0.0	8	0.75	4

The results lead to the observation that the way inconsistencies due to the cyber exploits are handled would make an appreciable difference in performance especially at the Command Interpretation stage. Thus, further research is needed to identify mechanisms that improve the resilience of a decision making organization to cyber exploits.

CONCLUSIONS

A new approach based on multi-formalism modeling to model, analyze, and evaluate the effect of cyber exploits on the coordination in organizations has been presented. The focus is on the effect that cyber exploits have on the performance of a decision making organization when its ability to share uncorrupted information in a timely manner is degraded due to cyber exploits on the networks that support the interactions between organization members. Two new measures of performance, information consistency and synchronization, were used to demonstrate the effect of cyber exploits, as well as the traditional accuracy and timeliness. The approach is now being applied to much larger examples in which different information sources, local and network-wide cyber exploits, and complex protocols can be tested.

REFERENCES

- Abu Jbara, A., A. H. Levis, and A. K. Zaidi, 2013. "On Using Multiple Interoperating Models to Address Complex Problems," in *Proc. Conference on Systems Engineering Research*, CSER 2013, (Atlanta, GA, March).
- CPN Tools, 2014: www.cpntools.org
- Grevet, J. L. and A. H. Levis, 1988. "Coordination in Organizations with Decision Support Systems," in *Proc. 1988 Symposium on C2 Research*, (Monterey, CA, June 7-9), SAIC, McLean, VA, 387-399.
- Gribaudo, M. and M. Iacono, 2014. "An Introduction to Multiformalism Modeling," in *Theory and Application of Multi-Formalism Modeling*, M. Gribaudo and M. Iacono (Eds.), IGI Global, Hershey, PA, 1-16.
- Hemingway, G., H. Neema, H. Nine, J. Sztipanovits, and G. Karsai, 2011. "Rapid Synthesis of High-Level Architecture-Based Heterogeneous Simulation: A Model-Based Integration Approach", *SIMULATION*, March.
- INET Framework, 2014: www.inet.omnetpp.org
- Jensen, K. and L. M. Kristensen, 2009. *Coloured Petri Nets*, Springer, Berlin.
- Levis, A. H., 1992. "A Colored Petri Net Model of Intelligent Nodes," in *Robotics and Flexible Manufacturing Systems*, J. C. Gentina and S. G. Tzafestas, (Eds.) Elsevier Science Publishers B. V., The Netherlands.
- Levis, A. H., A. K. Zaidi, and M. R. Rafi, 2012. "Multi-modeling and Meta-modeling of Human Organizations," in *Proc. 4th Int'l Conf. on Applied Human Factors and Ergonomics – AHFE2012* (San Francisco, CA, July).
- OMNeT ++, 2014: www.omnetpp.org
- Remy, P. A., A. H. Levis and V. Y. Jin, 1988. "On the Design of Distributed Organizational Structures," *Automatica*, Vol. 24, No. 1, pp. 81-86

ACKNOWLEDGMENT

This work was supported by the US Army Research Office through Contract No. W911NF-13-1-0154 to Carnegie Mellon University and sub-award No.1130163-311619 to George Mason University.

AUTHOR BIOGRAPHIES



ALEXANDER H. LEVIS is University Professor of Electrical, Computer, and Systems Engineering and heads the System Architectures Laboratory at George Mason University, Fairfax, VA. From 2001 to 2004 he served as the Chief Scientist of the U.S. Air Force. He was educated at Ripon College where he received the AB degree (1963) in Mathematics and Physics and then at MIT where he received the BS (1963), MS (1965), ME (1967), and Sc.D. (1968) degrees in Mechanical Engineering with control systems as his area of specialization. For the last fifteen years, his areas of research have been architecture design and evaluation, resilient architectures for command and control, and adversary multi-modeling for behavioral analysis. Dr. Levis is a Life Fellow of the Institute of Electrical and Electronic Engineers (IEEE) and past president of the IEEE Control Systems Society; a Fellow of the American Association for the Advancement of Science and of the International Council on Systems Engineering, and an Associate Fellow of the American Institute of Aeronautics and Astronautics. He has over 270 publications documenting his research, including the three volume set that he co-edited on *The Science of Command and Control*, and *The Limitless Sky: Air Force Science and Technology contributions to the Nation* published in 2004 by the Air Force.



Bahram Yousefi received the BS (2005), and MS (2008) degrees in electrical engineering from Azad University in Tehran, Iran. He also received a MS degree in computer engineering from George Mason University in 2012. He is currently working toward the PhD degree in the Department of Systems Engineering and Operations Research at the George Mason University (GMU). His current research mainly focuses on Model Based Systems Engineering (MBSE), multi-modeling, systems architecture evaluation, and architecture's resiliency. He is a member of INCOSE.

**Probability and Statistical
Methods for Modelling and
Simulation of High
Performance Information
Systems**

-

Special Session

ANALYSIS OF A FCFS QUEUE WITH TWO TYPES OF CUSTOMERS AND ORDER-DEPENDENT SERVICE TIMES

Bert Réveil, Dieter Claeys, Tom Maertens, Joris Walraevens, Herwig Bruneel
SMACS Research Group, TELIN Department
Ghent University
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
Email: {breveil, dclaeys, tmaerten, jw, hb}@telin.ugent.be

KEYWORDS

Queueing; Order-dependent service times; Class clustering

ABSTRACT

In this paper, we study a discrete-time first-come-first-served queueing system with a single server and two types (classes) of customers, where the (average) service time of a customer is longer if its type differs from the type of the preceding customer. As opposed to traditional literature, the different types of customers do not occur randomly and independently in the arrival stream: we include a Markovian type of correlation in the types of consecutive customers instead. We deduce the probability generating function of the system content, from which we extract various performance measures, such as the mean values of the system content and the customer delay. We demonstrate that the interclass correlation in the arrival stream has a tremendous impact on the system performance, which highlights the necessity to include it in the performance assessment of the system.

I. INTRODUCTION

In this paper, we study a discrete-time queueing system with two types (classes) of customers, a common queue, and one server. Customers are served in order of arrival, i.e., the queueing discipline is first-come-first-served (FCFS), irrespective of the class the consecutive customers belong to (referred to as “global FCFS” in this paper). However, the service time of a customer depends on the identity or non-identity of its class and the class of the preceding customer. Specifically, we assume that the (average) service time of any customer is longer if its type differs from the type of the previously served customer, a feature that arises regularly in practice. One major reason for this phenomenon may be the necessity to reconfigure or adapt the service facility for other tasks than the current one. Our model also applies to many other situations. Customers of distinct types could correspond, for instance, to vehicles that are heading to other destinations at road intersec-

tions, jobs with different execution times (because they require other resources), people requiring distinct kinds of services at a call center, (semi-finished) products that need different machines to be processed, printing jobs in a specialized printing house that delivers print work in several different formats, different types of goods being delivered in warehouses and being stocked in different sections, etc. The common feature of all these applications is that the service time of the next customer is, on average, longer if the preceding and the next customer (vehicle, job, person, product, printing job, goods delivery) require a different kind of service, i.e., do not belong to the same (service) class.

In traditional research on multi-class queues (see e.g. [1], [3], [8], [9], [13], [14], [18], [19]) it is standard to assume that the different types of customers occur randomly and independently in the arrival stream of customers into the system, which is often in contrast to the actual situation. In reality, there is usually some degree of *interclass correlation* or *class clustering*. In some cases, for instance, customers of the same type have a tendency to arrive “back to back”. As an example, consider a network router transmitting data from and towards various communicating processes running in the network. Within certain time frames of its service, it is likely that the router will transfer consecutive data packets that all originate from the same process.

In a number of recent papers [6], [7], [16], [15], [5], we have revealed that class clustering can have a major impact on the performance of several other queueing systems with two types of customers, such as queues with multiple class-dedicated servers and global FCFS service [6], [7], [16], priority queues [15], and single-server queues with global FCFS and class-dependent service times (regardless of the type of the previous customer) [5]. Therefore, we presume that this will also hold in the queueing system under investigation in the present paper, i.e., a system where the mean service time of a customer is longer when its type differs from the type of the previously served customer.

The paper is structured as follows. First, we describe the system in section II. Then, in section III,

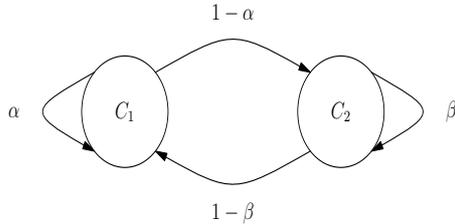


Fig. 1. Two-state Markov chain of the customer types.

we analyze the system behavior and we establish the probability generating function (PGF) of the number of customers in the system, hereafter referred to as the *system content*, both at customer departure times and at random slot boundaries. Next, the influence of class clustering is investigated in section IV, and finally, some conclusions are drawn and indications for further research given in section V.

II. SYSTEM DESCRIPTION

We study a discrete-time queueing system with an infinite waiting room, one server, and two types (classes) of customers, named C_1 and C_2 . The time axis is divided into fixed-length time intervals, referred to as *slots* in the sequel. New customers can arrive in the system at any given (continuous) point on the time axis, but customer service times can only start and end at slot boundaries. Customers are served according to a *global FCFS* service discipline, meaning that they are served in order of arrival, regardless of the class they belong to.

The arrival process of new customers is characterized in two steps. First, the total (aggregated) number of customer arrivals in consecutive slots is represented by a sequence of independent and identically distributed (IID) nonnegative discrete random variables with common probability mass function (PMF) $e(n)$ and probability generating function (PGF) $E(z)$:

$$e(n) \triangleq \text{Prob}[n \text{ arrivals in one slot}] , n \geq 0$$

$$E(z) \triangleq \sum_{n=0}^{\infty} e(n)z^n.$$

The (total) mean arrival rate, i.e., the (total) mean number of arrivals per slot, is given by

$$\lambda \triangleq E'(1). \tag{1}$$

Secondly, the occurrence of type C_1 and type C_2 customers within the total arrival stream is governed by a customer-type correlation model. This implies that we account for the possibility of *interclass correlation*, or *class clustering* in the arrival process. Customers of any given type may (or may not) have a tendency to “arrive back-to-back”. Consequently, the types of consecutive customers may be non-independent. In this study, we consider a first-order Markovian type of correlation between the types of consecutive customers (see Fig. 1).

If t_k denotes the type of customer k , the transition probabilities of the Markov chain that determines the

types of consecutive customers are defined as

$$\alpha \triangleq \text{Prob}[t_{k+1} = C_1 | t_k = C_1] ,$$

$$\beta \triangleq \text{Prob}[t_{k+1} = C_2 | t_k = C_2] . \tag{2}$$

The steady-state probabilities t_{C_1} and t_{C_2} of finding the Markov chain in state C_1 respectively C_2 are given by [10], [12]

$$t_{C_1} \triangleq \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_1] = \frac{1 - \beta}{2 - \alpha - \beta} ,$$

$$t_{C_2} \triangleq \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_2] = \frac{1 - \alpha}{2 - \alpha - \beta} . \tag{3}$$

They can be interpreted as the fractions of type C_1 and type C_2 customers in the arrival stream. Defining T_k as a numerical variable obeying

$$T_k = 1 \iff t_k = C_1 , \text{ and } T_k = 0 \iff t_k = C_2 ,$$

the steady-state correlation coefficient γ ($-1 \leq \gamma \leq 1$) of the Markov chain, called the *interclass correlation* in the sequel, is defined as

$$\gamma \triangleq \lim_{k \rightarrow \infty} \frac{E[T_k T_{k+1}] - E[T_k] E[T_{k+1}]}{\sqrt{\text{var}[T_k] \text{var}[T_{k+1}]}}$$

$$= \alpha + \beta - 1 . \tag{4}$$

It represents the amount of correlation between the types of two consecutive customers in the arrival stream (in the steady-state). Positive values of γ correspond to situations in which at least one customer type has a tendency to cluster. Negative values of γ typically imply (strongly) alternating customer type arrivals. If $\gamma = 0$, and consequently $\alpha = 1 - \beta$, the types of consecutive customers are independent, corresponding to the situation that is traditionally (implicitly) assumed in literature.

The *service time* of a customer indicates the number of slots needed to fully serve that customer. We assume that the service time of a customer depends on its own type and on the type of the previous customer. If both types are the same, the service time equals one slot, whereas in the other case, the service time is strictly positive and characterized by the PMF $b(n)$ ($n \geq 1$), PGF

$$B(z) \triangleq \sum_{n=1}^{\infty} b(n)z^n ,$$

and mean value

$$\mu_B^{-1} \triangleq B'(1) > 1 . \tag{5}$$

III. SYSTEM ANALYSIS

In this section, we first present an analysis of the total number of customers in the system at customer departure times. The PGF is established under steady-state conditions and a method is described to determine the two remaining unknowns in the expression. Finally, we deduce the PGF and the average value of the system content at random slot boundaries.

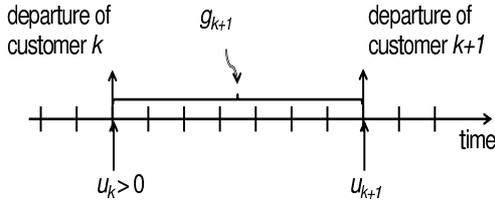


Fig. 2. Relationship between u_k and u_{k+1} when $u_k > 0$.

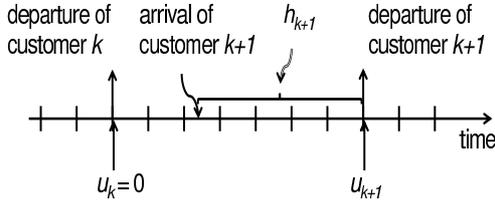


Fig. 3. Relationship between u_k and u_{k+1} when $u_k = 0$.

System equations at customer departure times

In this subsection, we establish system equations that capture the behavior of the system content at customer departure times. To this end, let u_k represent the total number of customers in the system immediately after the service completion of the k -th customer. Due to the assumptions presented in section II, the sequence of couples (t_k, u_k) constitutes a first-order Markov chain describing the evolution of the system from slot to slot. As described above, the state transitions for the sequence $\{t_k\}$ are governed by Equation (2). For the quantities $\{u_k\}$, we obtain two recursive equations that cover the complete set of situations depicted in figures 2 and 3:

$$\begin{aligned} u_{k+1} &= u_k - 1 + g_{k+1}, \text{ if } u_k > 0, \\ u_{k+1} &= h_{k+1}, \text{ if } u_k = 0. \end{aligned} \quad (6)$$

In these equations, g_{k+1} stands for the (total) number of arrivals during the service time of customer $k + 1$. The quantity h_{k+1} can be written as

$$h_{k+1} = g_{k+1} + f_{k+1},$$

with f_{k+1} the number of customer arrivals in the arrival slot of customer $k + 1$, but *after* customer $k + 1$ (in case customer $k + 1$ enters an empty system).

Its PMF $f(n)$ and PGF $F(z)$ can be found as

$$\begin{aligned} f(n) &\triangleq \text{Prob}[n \text{ additional arrivals} | \text{at least 1 arrival}] \\ &= \frac{e(n+1)}{1 - E(0)}, n \geq 0, \\ F(z) &\triangleq E[z^{f_{k+1}}] = \sum_{n=0}^{\infty} f(n)z^n \\ &= \frac{E(z) - E(0)}{z[1 - E(0)]}, \end{aligned} \quad (7)$$

irrespective of whether customer $k + 1$ is of the same or different type as customer k . For the PGFs of the quantities g_{k+1} and h_{k+1} , the equality of the customer types of two consecutive customers does make a difference. Taking into account that we are considering an

IID aggregated arrival process, implying that f_{k+1} and g_{k+1} are mutually independent, we find that

$$\begin{aligned} E[z^{g_{k+1}} | t_{k+1} = t_k] &= E(z), \\ E[z^{h_{k+1}} | t_{k+1} = t_k] &= F(z)E(z), \\ E[z^{g_{k+1}} | t_{k+1} \neq t_k] &= B(E(z)), \\ E[z^{h_{k+1}} | t_{k+1} \neq t_k] &= F(z)B(E(z)). \end{aligned}$$

System content at customer departure times

One of our intentions is to provide expressions for the performance measures of the queueing system under steady-state conditions. It is well-known [4], [17] that for any work-conserving queueing system, stability is guaranteed when the average amount of work entering the system per slot (often referred to as the *work load* ρ) is strictly less than the amount of work that can be delivered by the server per slot. In our model, considering a single server without interruptions, the stability condition thus boils down to

$$\rho \triangleq \lambda E[s] < 1,$$

with s the steady-state service time of an arbitrary customer. Using the law of the total expectation, $E[s]$ can be expanded, yielding

$$E[s] = t_A + t_B \mu_B^{-1}, \quad (8)$$

where $t_A = \alpha t_{C_1} + \beta t_{C_2}$ and $t_B = (1 - \alpha)t_{C_1} + (1 - \beta)t_{C_2}$ denote the steady-state probabilities that two consecutive customers belong to the same or the opposite class respectively. If we rework Equation (8), substituting α and β in terms of γ , t_{C_1} and t_{C_2} , another, more interesting expression for ρ can be found that links the work load directly to the amount of interclass correlation in the arrival process:

$$\rho = \lambda[1 + 2(1 - \gamma)(\mu_B^{-1} - 1)t_{C_1}t_{C_2}]. \quad (9)$$

As could have been anticipated, we find that in case of ultimate positive customer type correlation ($\gamma = 1$, i.e., both $\alpha = 1$ and $\beta = 1$), the work load reduces to λ . More generally, this also holds for single-class systems where either α or β is equal to 1, because in that case either t_{C_1} or t_{C_2} equals 0. If γ equals -1, i.e., if the customer type changes with every customer arrival, t_{C_1} and t_{C_2} are both equal to 0.5, and the work load increases to $\lambda \mu_B^{-1}$, also as expected.

Assuming that the stability condition is met, we define joint steady-state probabilities for the Markov chain $\{(t_k, u_k)\}$ as

$$\begin{aligned} p_{C_1}(i) &\triangleq \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_1, u_k = i], \\ p_{C_2}(i) &\triangleq \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_2, u_k = i], \end{aligned}$$

for all $i \geq 0$. The corresponding partial PGFs are given by

$$P_{C_1}(z) \triangleq \sum_{i=0}^{\infty} p_{C_1}(i)z^i, \quad P_{C_2}(z) \triangleq \sum_{i=0}^{\infty} p_{C_2}(i)z^i,$$

while the steady-state PGF $P(z)$ of the total system content at customer departure times is given by

$$P(z) = P_{C_1}(z) + P_{C_2}(z). \quad (10)$$

Relying on the balance equations of the Markov chain, it is now possible to establish two linear independent equations for the partial PGFs $P_{C_1}(z)$ and $P_{C_2}(z)$. For customers of class C_1 , we get

$$\begin{aligned} p_{C_1}(j) &= \lim_{k \rightarrow \infty} \text{Prob}[t_{k+1} = C_1, u_{k+1} = j] \\ &= \sum_{i=0}^{\infty} \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_1, u_k = i] \\ &\quad \text{Prob}[t_{k+1} = C_1, u_{k+1} = j | t_k = C_1, u_k = i] \\ &\quad + \sum_{i=0}^{\infty} \lim_{k \rightarrow \infty} \text{Prob}[t_k = C_2, u_k = i] \\ &\quad \text{Prob}[t_{k+1} = C_1, u_{k+1} = j | t_k = C_2, u_k = i] \\ &= \alpha \sum_{i=0}^{\infty} p_{C_1}(i) \lim_{k \rightarrow \infty} \text{Prob}[u_{k+1} = j | u_k = i, t_k = C_1, t_{k+1} = C_1] \\ &\quad + (1 - \beta) \sum_{i=0}^{\infty} p_{C_2}(i) \lim_{k \rightarrow \infty} \text{Prob}[u_{k+1} = j | u_k = i, t_k = C_2, t_{k+1} = C_1]. \end{aligned} \quad (11)$$

Taking the z-transform of (11) yields:

$$\begin{aligned} P_{C_1}(z) &\triangleq \sum_{j=0}^{\infty} p_{C_1}(j) z^j \\ &= \alpha \sum_{i=0}^{\infty} p_{C_1}(i) \lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = i, t_k = C_1, t_{k+1} = C_1] \\ &\quad + (1 - \beta) \sum_{i=0}^{\infty} p_{C_2}(i) \lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = i, t_k = C_2, t_{k+1} = C_1]. \end{aligned} \quad (12)$$

The expectations in the above equations can be elaborated using the system equations in (6):

$$\begin{aligned} &\lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = i, t_k = C_1, t_{k+1} = C_1] \\ &= \lim_{k \rightarrow \infty} \text{E}[z^{i-1+g_{k+1}} | t_k = C_1, t_{k+1} = C_1] \\ &= z^{i-1} E(z), \text{ for all } i \geq 1, \end{aligned} \quad (13)$$

$$\begin{aligned} &\lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = 0, t_k = C_1, t_{k+1} = C_1] \\ &= \lim_{k \rightarrow \infty} \text{E}[z^{h_{k+1}} | t_k = C_1, t_{k+1} = C_1] \\ &= F(z)E(z), \text{ for } i = 0, \end{aligned} \quad (14)$$

$$\begin{aligned} &\lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = i, t_k = C_2, t_{k+1} = C_1] \\ &= \lim_{k \rightarrow \infty} \text{E}[z^{i-1+g_{k+1}} | t_k = C_2, t_{k+1} = C_1] \\ &= z^{i-1} B(E(z)), \text{ for all } i \geq 1, \end{aligned} \quad (15)$$

$$\begin{aligned} &\lim_{k \rightarrow \infty} \text{E}[z^{u_{k+1}} | u_k = 0, t_k = C_2, t_{k+1} = C_1] \\ &= \lim_{k \rightarrow \infty} \text{E}[z^{h_{k+1}} | t_k = C_2, t_{k+1} = C_1] \\ &= F(z)B(E(z)), \text{ for } i = 0. \end{aligned} \quad (16)$$

Substitution of (13), (14), (15) and (16) in expression (12) finally leads to a first linear equation between

$P_{C_1}(z)$ and $P_{C_2}(z)$:

$$\begin{aligned} (z - \alpha E(z))P_{C_1}(z) &= (1 - \beta)B(E(z))P_{C_2}(z) \\ &\quad + \alpha E(z)(zF(z) - 1)P_{C_1}(0) \\ &\quad + (1 - \beta)B(E(z))(zF(z) - 1)P_{C_2}(0). \end{aligned} \quad (17)$$

Similarly, a second linear equation can be found starting from the balance equations for type C_2 customers:

$$\begin{aligned} (z - \beta E(z))P_{C_2}(z) &= (1 - \alpha)B(E(z))P_{C_1}(z) \\ &\quad + \beta E(z)(zF(z) - 1)P_{C_2}(0) \\ &\quad + (1 - \alpha)B(E(z))(zF(z) - 1)P_{C_1}(0). \end{aligned} \quad (18)$$

Equations (17) and (18) can be solved for the unknown partial PGFs $P_{C_1}(z)$ and $P_{C_2}(z)$. Using the results and Equation (7) to expand Equation (10), we obtain a first expression for the PGF $P(z)$:

$$\begin{aligned} P(z) &= \frac{P(0)(E(z) - 1)}{1 - E(0)} \times \\ &\quad \frac{z(p_A E(z) + p_B B(E(z))) - \alpha \beta E(z)^2 + (1 - \alpha)(1 - \beta)B(E(z))^2}{z^2 - z(\alpha + \beta)E(z) + \alpha \beta E(z)^2 - (1 - \alpha)(1 - \beta)B(E(z))^2}, \end{aligned} \quad (19)$$

with

$$\begin{aligned} p_A &\triangleq \frac{\alpha P_{C_1}(0) + \beta P_{C_2}(0)}{P(0)}, \\ p_B &\triangleq \frac{(1 - \alpha)P_{C_1}(0) + (1 - \beta)P_{C_2}(0)}{P(0)}. \end{aligned} \quad (20)$$

Given that $P(0) = P_{C_1}(0) + P_{C_2}(0)$, these quantities are equal to

$$\begin{aligned} p_A &= \lim_{k \rightarrow \infty} \text{Prob}[t_{k+1} = t_k | u_k = 0], \\ p_B &= \lim_{k \rightarrow \infty} \text{Prob}[t_{k+1} \neq t_k | u_k = 0], \end{aligned} \quad (21)$$

the conditional probabilities that a new customer entering an empty system (in steady state) is of the same or the opposite type respectively as the last customer that was served by the system.

Expression (19) still contains three unknowns that need to be determined: $P(0)$, p_A and p_B . The probability $P(0)$ can be found by imposing the normalization condition on the PGF $P(z)$, i.e. $P(1) = 1$. Using de l'Hôpital's rule to solve the equation, we obtain that

$$P(0) = \frac{(1 - E(0))(1 - \rho)}{\lambda}. \quad (22)$$

In order to derive expressions for p_A and p_B , two linear equations in p_A and p_B are established. The first one simply states that

$$p_A + p_B = 1. \quad (23)$$

The second equation follows from the fact that a PGF, like $P(z)$, is bounded inside the closed unit disk of the complex z-plane $\{z \in \mathbb{C} : |z| \leq 1\}$. As it can be proved via Rouché's theorem [2] that the denominator of $P(z)$ has exactly two zeroes inside the closed unit disk, the aforementioned property of PGFs implies that those zeroes, one of which is equal to 1, must also be zeroes of $P(z)$'s numerator. Otherwise, the function value would

tend to infinity inside the closed unit z -disk. For $z = 1$, given its factor $(E(z) - 1)$, the numerator clearly vanishes. For the second zero however, called \hat{z} from here on, the other factor in the numerator should equal 0, which yields a linear equation for p_A and p_B . Solving this equation, in combination with Equation (23), we find that p_A and p_B can be determined as

$$\begin{aligned} p_A &= \frac{(\alpha + \beta)E(\hat{z}) - B(E(\hat{z})) - \hat{z}}{E(\hat{z}) - B(E(\hat{z}))}, \\ p_B &= \frac{(1 - \alpha - \beta)E(\hat{z}) + \hat{z}}{E(\hat{z}) - B(E(\hat{z}))}. \end{aligned} \quad (24)$$

Once the zero \hat{z} is computed numerically, e.g. via standard root-finding functions in Maple or Matlab, p_A and p_B are fixed, and as such, so is $P(z)$.

System content at random slot boundaries

From earlier research [4], it follows that for all discrete-time single-server queueing systems with (general) independent customer arrivals from slot to slot (with PGF $E(z)$), a fairly simple relationship holds between the PGF $P(z)$ of the system content at customer departure times and the PGF $U(z)$ of the system content at random slot boundaries, regardless of the exact characteristics of the service process and the intra-slot details of the arrival process (e.g., single or batch arrivals, the exact time customers arrive within the slot, etc.). This relationship is

$$P(z) = \frac{E(z) - 1}{\lambda(z - 1)} U(z). \quad (25)$$

As the examined model belongs to the class of systems described above, relationship (25) in combination with Equations (19) and (22) leads to the following expression for the PGF of the system content at random slot boundaries:

$$U(z) = (1 - \rho)(z - 1) \times \frac{z(p_A E(z) + p_B B(E(z))) - \alpha\beta E(z)^2 + (1 - \alpha)(1 - \beta)B(E(z))^2}{z^2 - z(\alpha + \beta)E(z) + \alpha\beta E(z)^2 - (1 - \alpha)(1 - \beta)B(E(z))^2}. \quad (26)$$

From this expression, various interesting performance measures can be derived, one of which is the mean system content $E[u]$ at random slot boundaries. The latter can be determined as $E[u] = U'(1)$, where u represents the system content at the beginning of a random slot in steady state. After long and tedious calculations, we find that

$$\begin{aligned} E[u] &= \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1 - \rho)} \\ &+ 2\lambda(1 - \mu_B^{-1})t_{C_1}t_{C_2} + \frac{p_B\lambda(\mu_B^{-1} - 1)}{1 - \gamma} \\ &+ \frac{(1 - \gamma)\lambda^2(\mu_B^{-1} - 1)^2 t_{C_1}t_{C_2}(1 - 4t_{C_1}t_{C_2})}{1 - \rho}. \end{aligned} \quad (27)$$

with $C'(1)$ and $C''(1)$ the first two derivatives for $z = 1$ of the PGF $C(z)$ of the service time of an arbitrary customer:

$$C(z) = t_A z + (1 - t_A)B(z), \quad (28)$$

with

$$t_A = \lim_{k \rightarrow \infty} \text{Prob}[t_k = t_{k-1}].$$

In Equation (27), the first term ρ accounts for the average server content, or the mean number of customers in service. The last four terms cover the mean queue occupancy, meaning the average number of customers that are waiting to be served.

Higher-order moments of the system content at random slot boundaries can be obtained by computing higher-order derivatives of the PGF $U(z)$. By means of Little's law (for discrete-time queues) [11], one can determine the average *delay* (system time) of an arbitrary customer as $E[d] = E[u] / \lambda$ (d stands for the delay of a random customer in the system in steady state). The mean of the *waiting time* w is obtained as $E[w] = E[d] - E[s]$, where $E[s]$ was defined in (8). In our case, it is given by

$$\begin{aligned} E[w] &= \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2\lambda(1 - \rho)} \\ &+ 2(1 - \mu_B^{-1})t_{C_1}t_{C_2} + \frac{p_B(\mu_B^{-1} - 1)}{1 - \gamma} \\ &+ \frac{(1 - \gamma)\lambda(\mu_B^{-1} - 1)^2 t_{C_1}t_{C_2}(1 - 4t_{C_1}t_{C_2})}{1 - \rho}. \end{aligned}$$

IV. DISCUSSION OF RESULTS AND NUMERICAL EXAMPLES

In this section, we discuss the obtained results, both from a qualitative perspective as by means of some numerical examples. The first interesting result was already given by Equation (9). The equation expresses the direct dependency of the work load ρ on the inter-class correlation factor γ ($\triangleq \lambda E[s]$). Consequently, the stability condition,

$$\lambda < \frac{1}{E[s]} = \frac{1}{1 + 2(1 - \gamma)(\mu_B^{-1} - 1)t_{C_1}t_{C_2}}, \quad (29)$$

reveals that the supremum of the achievable throughput of the presented system, denoted as λ_{sup} , and expressed in customers per slot, depends on γ , μ_B^{-1} and the fractions of C_1 and C_2 customers.

Equation (29) reveals that if μ_B^{-1} increases, λ_{sup} decreases, because the mean service time for customers following customers of the opposite type is increased.

For fixed μ_B^{-1} , we find that λ_{sup} is lowest when $t_{C_1}t_{C_2}$ reaches its maximal value, i.e., for $t_{C_1} = t_{C_2} = \frac{1}{2}$. If one type of customers enters the system more often than the other ($t_{C_1} > 0.5 > t_{C_2}$ or $t_{C_2} > 0.5 > t_{C_1}$), consecutive customers will be of the same type more often, implying that the average service time of an arbitrary customer decreases, or that the throughput of the system increases.

When t_{C_1} , t_{C_2} and μ_B^{-1} are fixed, the throughput decreases when the types of consecutive customers begin to alter more regularly, i.e., when γ becomes smaller. The worst case occurs when $\gamma = -1$ meaning that the types of subsequent customers always differ. The best scenario occurs when only one type of customers enters the system ($\gamma = 1$).

A second interesting result was given by Equation (27). This expression clearly indicates the influence of the different system parameters on the mean system content at random slot boundaries. The first two terms of Equation (27) correspond to the classical terms that constitute the expression for the average system content at random slot boundaries of a system with no interclass correlation and a service-time PGF $C(z)$. The other three terms in the expression can be fully attributed to the presence of class clustering in the arrival process.

It is not surprising to see that the mean system content depends on the first two moments of the aggregated arrival process (represented by the quantities λ , $E''(1)$ and $\rho = \lambda E[s]$) and on the first two moments of the service times (represented by the quantities $C'(1)$, $C''(1)$, μ_B^{-1} and $\rho = \lambda C'(1)$). Furthermore, we anticipated to find that the mean system content goes to infinity as soon as the work load ρ approaches its limiting value 1.

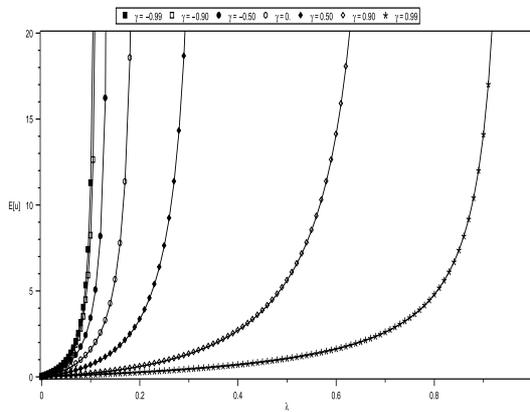


Fig. 4. Mean system content versus the mean arrival rate λ , for Poisson arrivals, $B(z)$ given by (30), $\mu_B^{-1} = 9$, $t_{C_1} = t_{C_2} = 0.5$ and several interclass correlation factors.

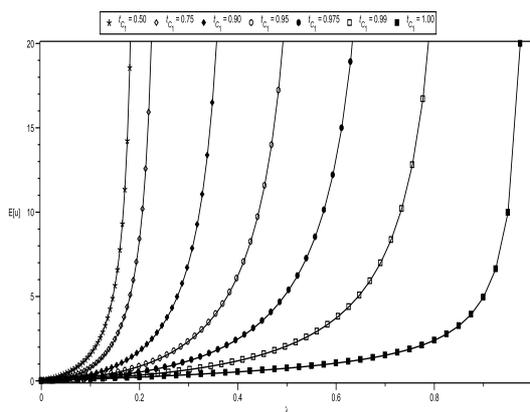


Fig. 5. Mean system content versus the mean arrival rate λ , for Poisson arrivals, $B(z)$ given by (30), $\mu_B^{-1} = 9$, $\gamma = 0$ and various fractions of customer types in the arrival stream.

In Figures 4-6, we present numerical results for two-class queueing systems dealing with an aggregated Poisson arrival process (i.e., $E(z) = e^{\lambda(z-1)}$) and the following PGF of the service time of customers following

a customer of the opposite type:

$$B(z) = \frac{z}{\mu_B^{-1} + (1 - \mu_B^{-1})z} . \quad (30)$$

Figure 4 shows the mean system content versus λ for different values of γ , in a system where $\mu_B^{-1} = 9$ and both types of customers occur with the same a priori frequency (i.e., $t_{C_1} = t_{C_2} = 0.5$). The average number of customers the system can deal with depends heavily on the amount of interclass correlation: the more positive, the more customers can be served per slot. This implies that the system occupancy raises rapidly for systems with a negative interclass correlation.

In Figure 5, we examine the impact of the fractions of type C_1 and type C_2 customers in the arrival stream on the average system content. The figure depicts the mean system content versus λ , in a system where $\mu_B^{-1} = 9$, and with a fixed interclass correlation of 0.

The figure mainly shows that having two types of customers instead of one, strongly affects the mean system content. If only one type of customer occurs, the average system content is much lower, because every arriving customer only requires one time slot to be served. As soon as two different types of customers enter the system, the average system content increases considerably. As reasoned before, based on Equation (29), the exact fraction of type C_1 and type C_2 customers influences the achievable throughput of the system.

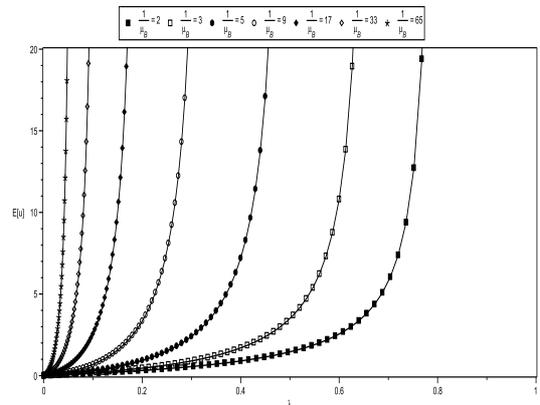


Fig. 6. Mean system content versus the mean arrival rate λ , for Poisson arrivals, $\gamma = 0.5$ and $t_{C_1} = t_{C_2} = 0.5$; $B(z)$ given by (30).

In a third plot (Figure 6), we present the mean system content of a system that is facing a positive interclass correlation of 0.5 and an equal amount of type C_1 and type C_2 customers. The mean service time μ_B^{-1} is varied. The average system content increases when μ_B^{-1} increases. If the interclass correlation factor is fixed, a longer service time for customers that are not of the same type as the previous customer implies more arriving customers during that service time, and consequently, more customers waiting in the system.

V. CONCLUSIONS AND FURTHER RESEARCH

We have investigated a discrete-time queueing system with two types (classes) of customers, one server,

a common queue, global FCFS queueing discipline, and service times that are, on average, longer when the types of consecutive customers differ. The aim of the paper was to study the impact of class clustering on the system performance, a feature that is traditionally overlooked in the literature. As we have already revealed that class clustering has a major impact on other systems, we presumed that this statement also holds for a system with order-dependent service times. We have therefore studied the performance, both in terms of stability and system content, in such a system that is subject to class clustering. We have found that class clustering has an undeniable impact on the system performance. This conclusion highlights the necessity to incorporate class clustering when studying a queue with order-dependent service times. Our results can be used for that purpose.

This paper is a first (conceptual) step in a more general study of queueing systems with order-dependent service times. A first extension could be a model in which the service-time distribution of a customer is not only dependent on the identity or non-identity of its type and the previous customer's type, but also on the actual type itself. In the present paper, the service times are deterministic and equal to one slot when the customer at hand has the same type as the preceding customer, while the service time is, on average, longer than one slot in the opposite case; this could be generalized such that in both cases the service-time distributions are different and completely arbitrary. Future research may also consider non-Markovian types of interclass correlation in the arrival stream, e.g., arrival processes where the numbers of consecutive customers of the same type have more general probability distributions than the geometric distributions considered in this paper. Another interesting extension is to incorporate slot-to-slot correlation in the (aggregated) arrival process, both with respect to the total numbers of arrivals per slot and the types of customers arriving in different slots. A generalization to more customer types seems complicated, especially when the state transitions of the modulating Markov chain of the arrival process are completely arbitrary. A matrix-analytic approach may be more feasible here, but the derivation of explicit closed-form results could be harder. Finally, it may be interesting to investigate related continuous-time queueing models. Some of these issues will be dealt with in the future.

REFERENCES

[1] I.J.B.F. Adan, A. Sleptchenko, and G.J. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research*, 26(4):559–585, 2009.

[2] I.J.B.F. Adan, J.S.H. van Leeuwen, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[3] M.A.A. Boon, I.J.B.F. Adan, and O.J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67:468–484, 2010.

[4] H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.

[5] H. Bruneel, T. Maertens, B. Steyaert, D. Claeys, D. Fiems, and J. Walraevens. Analysis of a two-class FCFS queueing system with interclass correlation. In *Proceedings of the 19th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA'12)*, Grenoble, June 4-6 2012.

[6] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. Impact of blocking when customers of different classes are accommodated in one common queue. In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems (ICORES)*, Villamoura, Portugal, February 2012.

[7] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *European Journal of Operational Research*, 223:123–132, 2012.

[8] H. Chen and H. Zhang. Stability of multiclass queueing networks under priority service disciplines. *Operations Research*, 48(1):26–37, 2000.

[9] E.B. Cil, F. Karaesmen, and E.L. Ormeci. Dynamic pricing and scheduling in a multi-class single-server queueing system. *Queueing Systems*, 67(4):305–331, 2011.

[10] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, Third Edition*. Wiley, New York, 1968.

[11] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.

[12] R.G. Gallager. *Discrete stochastic processes*. Kluwer Academic, Boston/Dordrecht/London, 1996.

[13] D. Gamarnik and D. Katz. On deciding stability of multiclass queueing networks under buffer priority scheduling policies. *Annals of Applied Probability*, 19(5):2008–2037, 2009.

[14] M. Larrañaga, U. Ayesta, and I.M. Verloop. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation*, 70(10):841–858, 2013.

[15] T. Maertens, H. Bruneel, and J. Walraevens. Effect of class clustering on delay differentiation in priority scheduling. *Electronic Letters*, 48(10):568–569, 2012.

[16] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys, and J. Walraevens. Impact of class clustering and global FCFS service discipline on the system occupancy of a two-class queueing model with two dedicated servers. In *Proceedings of the 7th International Conference on Queueing Theory and Network Applications (QTNA 7)*, Kyoto, Japan, 2012.

[17] H. Takagi. *Queueing analysis - vol. 3: discrete-time systems*. North Holland, 1993.

[18] O.S. Ulusu and T. Altıok. Waiting time approximation in multi-class queueing systems with multiple types of class-dependent interruptions. *Annals of Operations Research*, 202(1):185–195, 2013.

[19] I.M. Verloop, U. Ayesta, and S. Borst. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems - Theory and Applications*, 20(4):473–509, 2010.

JOINT STATIONARY DISTRIBUTION OF QUEUES IN HOMOGENOUS $M|M|3$ QUEUE WITH RESEQUENCING

Ilaria Caraccio
University of Salerno
84084-Via Giovanni Paolo II, 132
Fisciano (SA) Italy
Email: icaraccio@unisa.it

Alexander V. Pechinkin
Rostislav V. Razumchik
Institute of Informatics Problems of RAS
Vavilova, 44-2,
119333, Moscow, Russia
Email: apechinkin@ipiran.ru,
rrazumchik@ieee.org

KEYWORDS

Resequencing, queueing system, joint distribution.

ABSTRACT

Resequencing issue is a crucial issue in simultaneous processing systems where the order of customers (jobs, units) upon arrival has to be preserved upon departure. In this paper stationary characteristics of $M/M/3/\infty$ queueing system with reordering buffer of infinite capacity are being analyzed. Noticing that customer in reordering buffer may form two separate queues, focus is given to the study of their size distribution. Expressions for joint stationary distribution are obtained both in explicit form and in terms of generating functions. Numerical example is presented.

INTRODUCTION

Resequencing issue is a crucial issue in simultaneous processing systems where the order of customers (jobs, units, et.c.) upon arrival has to be preserved upon departure. Various analytical methods and models have been proposed to study the impacts of resequencing. A general survey of queueing theoretic methods and early models for the modeling and analysis of parallel and distributed systems with resequencing can be found in Boxma et al. (1994). Survey on the resequencing problem that covers period up to 1997 can be found in Dimitrov (1997). Queueing-theoretic approach to resequencing problem implies that the system under consideration is represented as interconnected queueing systems/networks. Following Leung et al. (2010) existing papers can be grouped into categories: papers that characterize the disordering process through a queueing system with several servers sharing a single queue (see, e.g. Agrawal and Ramaswami (1987)) and papers where disordering is modeled by a queueing system with several parallel servers and queues, and each server has its own dedicated queue (see, e.g. Ye Xia et al. (2008)). For a short survey of these two categories see e.g. Leung et al. (2010). Following this approach various problems setting have been considered and solved including distribution of number of packets in reordering buffer and in

system under different assumptions about arrival and service process (see, e.g. Jain and Sharma (2011), Lelarge (2008), Chakravarthy (1998), Takine et al. (1994), De Nicola C. et al. (2013)); distribution and/or mean of the resequencing delay (see, e.g. Huisman and Boucherie (2002), Ding et al. (1991)), end-to-end (i.e. sender-receiver) delay (see, e.g. Chowdhury (1991)); large deviations of the queue size in reordering buffer (see, e.g. Gao et al. (2012)), asymptotics of the resequencing delay (see, e.g. Jun Li et al. (2010)), optimal allocation of customers to servers (Gogate and Panwar (1999), optimization issues (Dimitrov et al. (2002)). Among practical related papers one can also cite Leung et al. (2010), Zheng et al. (2010) and Wen-Fen (2011), Li et al. (2010), Huisman and Boucherie (2001), Min Choi et al. (2012), Rubin et al. (1991).

In this paper we propose new problem statement for systems with resequencing that are modeled by multiserver queues followed with infinite resequencing buffer. New problem is motivated by noticing that customers awaiting in resequencing buffer may form separate queues. The most convenient way to explain how queues are separated in resequencing buffer is with the example. Consider a queueing system with three servers, infinite capacity main buffer and reordering buffer. Let the state of the system at some instant be as depicted in Fig. 1. In squares one can see customers' sequential numbers. White (black) squares in Fig. 1 mean that customers with these sequential numbers have received (have not yet received) service. Here one can distinguish two queues: one which is formed by customers awaiting customer no. 18 (queue #1), another is formed by customers awaiting customer no. 15 (queue #2). Three cases need to be considered.

Case 1. Now on if customer no. 21 is next to complete its service then it joins queue #1 and stays there until service of customer no. 18 is complete. Customer no. 22 joins idle server.

Case 2. If customer no. 15 is next to complete its service then it goes through queue #1 without waiting and joins queue #2. Meanwhile customer no. 22 joins idle server. As there is no customer in the system with sequential number smaller than any sequential number in

queue #2, then all customers from queue #2 leave the system. Resequencing buffer “sees”, that queue #2 is empty and moves its contents to queue #2. Now there are three options. First: if customer no. 18 is next to complete service, then it goes through queue #1 without waiting and joins queue #2. Customer no. 23 joins idle server. Again there is no customer in the system with sequential number smaller than any sequential number in queue #2. Thus all customers from queue #2 leave the system. Resequencing buffer becomes empty. Now if customer no. 21 is next to complete service, it leaves the system. If customer no. 22 is next to complete service, it goes through queue #1 without waiting and joins queue #2 where it waits for customer no. 21. Finally, if customer no. 23 is next to complete service, it joins queue #1 and does not proceed to queue #1 because it needs customer no. 22 to complete its service before both of them may join queue #2. Second: if customer no. 21 is next to complete service, then it goes through queue #1 again without waiting, joins queue #2 and waits there with other customer for service completion of customer no. 18. Third: if customer no. 22 is next to complete service, then customer no. 23 joins idle server, customer no. 22 joins queue #1 and stops there because “sees” gap between its sequential number and largest sequential number in queue #2. It waits there for customer no. 21.

Case 3. If customer no. 18 is the first to complete its service then it joins queue #1 and customer no. 22 joins idle server. Resequencing buffer “sees”, that there is no gap in the middle of sequence and moves the content of queue #1 to queue #2 (queue #1 becomes empty). Now there are again three options. First: if customer no. 15 is next to complete service, then it goes through queue #1 without waiting, joins queue #2 and immediately (because the sequence is complete) leaves the system with all contents of queue #2. Second: if customer no. 21 is next to complete service, then it goes through queue #1 again without waiting, joins other customers in queue #2 that wait for service completion of customer no. 15. Third: if customer no. 22 is next to complete service, then it joins queue #1 and stops there, because “sees” gap between its sequential number and the largest sequence number in queue #2. The operation of the system proceeds along the line.

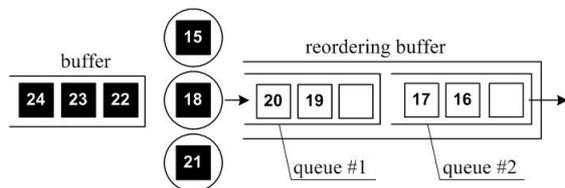


Figure 1: Scheme of the model

Clearly, when the number of server is n there are $(n-1)$ queues in resequencing buffer. Sum their contents is the total number of customers in resequencing buffer. The main contribution of this paper are algorithm and probability generating function of joint stationary prob-

abilities of the number of customers in buffer, queue #1 and queue #2.

The paper is organized as follows. In the next section we give detailed description of the system. Then we find joint stationary distribution both algorithm-wise and in terms of probability generating functions. The last section is devoted to numerical results.

DESCRIPTION OF THE SYSTEM

Consider a queueing system with three servers, infinite capacity buffer, incoming poisson flow of customers (of intensity λ) and exponential distribution of service time at each server (with parameter μ) and resequencing buffer (RB) of infinite capacity. Customers upon entering the system obtain sequential number and join buffer. Without loss of generality we suppose that the sequence starts from 1 and coincides with the row of natural numbers, i.e. the first customer upon entering the (empty) system receives number 1, the second — number 2 and so on and so forth. Customers leave the system strictly in order of their arrival (i.e. in the sequence order). Thus after customer’s arrival it remains in the buffer for some time and then receives service when one of the servers becomes idle. If at the moment of its service completion there are no customers in the system or all other customers present at that moment in the queue and the rest two servers have greater sequential numbers it leaves the system. Otherwise it occupies one place in the RB. Customer from RB leaves it if and only if its sequential number is less than sequential numbers of all other customers present in system. Thus customers may leave RB in groups.

Let us call “1st level” customer the one which is in service and was the last to enter server; “2nd level” customer is the one which is in service and was the penultimate to enter server; finally, “3rd level” the customer is the one which is in service and was the first to enter server. If the number of busy servers is 3, then customers that entered RB between “1st level” and “2nd level” customer form queue #1; customers which entered RB between “2nd level” and “3rd level” customer form queue #2. If the number of busy servers is 2, then customers which entered RB after “1st level” customer form queue #1; customers which entered RB between “1st level” and “2nd level” customer form queue #2. When there is only one busy server all customers in RB form queue #2.

The operation of the considered queueing system can be completely described by Markov process $\zeta(t) = \{(\xi(t), \eta(t), \nu(t)), t \geq 0\}$ with three components: $\xi(t)$ — number of customers in buffer and server at time t , $\eta(t)$ — number of customers in queue #1 of RB at time t , $\nu(t)$ — number of customers in queue #2 of RB at time t . In case $\xi(t) = 0$, the second and third component of $\zeta(t)$ are omitted; in case $\xi(t) = 1$, the second is omitted. The state space of $\zeta(t)$ is $\mathcal{X} = \{0\} \cup \{(1, i), i \geq 0\} \cup \{(n, i, j), n \geq 2, i \geq 0, j \geq 0\}$. Henceforth it is assumed that service and arrival processes are mutually independent and necessary and suf-

efficient condition of stationarity $\rho/3 < 1$, where $\rho = \lambda/\mu$ holds for the system.

STATIONARY JOINT PROBABILITY DISTRIBUTION

Note that the total number of customers in servers and buffer of the considered QS with resequencing coincides with the total number of customers in $M/M/3/\infty$ queue. Therefore, its stationary distribution $\{p_i, i \geq 0\}$, has the form

$$p_0 = \left(\sum_{i=0}^2 \frac{\rho^i}{i!} + \frac{\rho^3}{2!(3-\rho)} \right)^{-1}, \quad (1)$$

$$p_i = \frac{\rho^i}{i!} p_0, \quad i = \overline{1, 3}, \quad (2)$$

$$p_i = \frac{\rho^i}{3!3^{i-3}} p_0 = \tilde{\rho}^{i-3} p_3, \quad \tilde{\rho} = \frac{\rho}{3}, \quad i \geq 4. \quad (3)$$

Provided that RB is empty when servers are idle, p_0 is also the probability of the considered system with resequencing to be empty.

Denote by $p_{n;i,j}, n \geq 3, i \geq 0, j \geq 0$, stationary probability of the fact that there are n customers in servers and buffer, i customers in queue #1 of RB, j customers in queue #2 of RB. By $p_{n;i}, n \geq 3, i \geq 0$, denote stationary probability of the fact that there are n customers in servers and buffer and i customers in queue #1 of RB. Clearly $p_{n;i} = \sum_{j \geq 0} p_{n;i,j}$. Probabilities $p_{2;i,j}, i \geq 0, j \geq 0$ and $p_{2;i}, i \geq 0$, are defined by analogy. Finally, let $p_{1;i}, i \geq 0$, be stationary probability of the fact that there is only one busy server and i customers reside in queue #2 of RB. Note that distribution $p_n, n \geq 0$, of the total number of customers in servers and buffer (which is defined by (1)–(3)) can be expressed as follows

$$p_1 = \sum_{i \geq 0} p_{1;i}, \quad p_n = \sum_{i \geq 0} \sum_{j \geq 0} p_{n;i,j}, \quad n \geq 2.$$

In the next section we proceed to the derivation of the steady-state equations for the defined above probabilities.

SYSTEM OF EQUILIBRIUM EQUATIONS

The derivation of steady-state equations we start with $p_{n;i}, n \geq 3$, — stationary probabilities of the fact that there are total of n customers in servers and buffer and i customers in queue #1 of RB. The easiest way to do this is to use global balance principle. Let $i = 0$. The set of states, corresponding to probability $p_{n;0}$ when $n \geq 3$ is $\cup_{j \geq 0} (n, 0, j)$. Mean rate out of this set is clearly $(\lambda + 3\mu)p_{n;0}$. Note that one can enter the set $\cup_{j \geq 0} (n, 0, j)$ either from set of states $\cup_{j \geq 0} (n-1, 0, j)$ (with mean rate $\lambda p_{n-1;0}$) or from $\cup_{i \geq 0} \cup_{j \geq 0} (n+1, i, j)$ (with mean rate $3\mu \frac{2\mu}{3\mu} p_{n+1} = 2\mu p_{n+1}$). Putting it altogether yields

$$(\lambda + 3\mu)p_{n;0} = \lambda p_{n-1;0} + 2\mu p_{n+1}, \quad n \geq 3. \quad (4)$$

The derivation of other steady-state equations is done by analogy utilizing the properties of exponentially distributed random variables and thus omitted. Probabilities $p_{n;i}, n \geq 1$ satisfy

$$(\lambda + 3\mu)p_{n;i} = \lambda p_{n-1;i} + \mu p_{n+1;i-1}, \quad n \geq 3, \quad i \geq 1, \quad (5)$$

whereas for probabilities $p_{2;i}, i \geq 0$, it holds

$$(\lambda + 2\mu)p_{2;0} = \lambda p_1 + 2\mu p_3, \quad (6)$$

$$(\lambda + 2\mu)p_{2;i} = \mu p_{3;i-1}, \quad i \geq 1. \quad (7)$$

Probabilities $p_{1;i}, i \geq 0$, satisfy

$$(\lambda + \mu)p_{1;0} = \lambda p_0 + \mu p_{2;0}, \quad (8)$$

$$p_{1;i}(\lambda + \mu) = \mu p_{2;i} + \mu \sum_{j=0}^{i-1} p_{2;i-j-1,j}, \quad i \geq 1. \quad (9)$$

One can verify that for probabilities $p_{n;i,j}, n \geq 3, i \geq 0, j \geq 0$, the following equations hold:

$$(\lambda + 3\mu)p_{n;0,0} = \lambda p_{n-1;0,0} + \mu p_{n+1;0}, \quad n \geq 3, \quad (10)$$

$$(\lambda + 3\mu)p_{n;0,j} = \lambda p_{n-1;0,j} + \mu p_{n+1;j} +$$

$$\mu \sum_{k=0}^{j-1} p_{n+1;k,j-k-1}, \quad n \geq 3, \quad j \geq 1, \quad (11)$$

$$(\lambda + 3\mu)p_{n;i,j} = \lambda p_{n-1;i,j} +$$

$$\mu p_{n+1;i-1,j}, \quad n \geq 3, \quad i \geq 1, \quad j \geq 0. \quad (12)$$

Finally, probabilities $p_{2;i,j}, i \geq 0, j \geq 0$ satisfy

$$(\lambda + 2\mu)p_{2;0,0} = \lambda p_{1;0} + \mu p_{3;0}, \quad (13)$$

$$(\lambda + 2\mu)p_{2;0,j} = \lambda p_{1;j} + \mu p_{3;j} +$$

$$\mu \sum_{k=0}^{j-1} p_{3;k,j-k-1}, \quad j \geq 1, \quad (14)$$

$$(\lambda + 2\mu)p_{2;i,j} = \mu p_{3;i-1,j}, \quad i \geq 1, \quad j \geq 0. \quad (15)$$

The analysis of steady-state equations resulted in the development of simple algorithm for step-by-step computation of stationary joint probabilities $p_{n;i,j}, n \geq 2, i \geq 0, j \geq 0$, and $p_{n;i}, n \geq 1, i \geq 0$. The algorithm is given below.

For practical purposes it may be sometimes sufficient to know either only $\pi_{n;i}, n \geq 1, i \geq 0$, — stationary probabilities of the fact that total number of customers in servers and in buffer is n and total number of customers in RB (sum of queue #1 and queue #2) is i , or only $\pi_i, i \geq 0$, — stationary probabilities of the fact that there are n customers in total in the whole system (including buffer, servers, RB). These quantities can be calculated from joint probability distribution as follows

$$\pi_{1;i} = p_{1;i}, \quad i \geq 0, \quad \pi_{2;i} = \sum_{j=0}^i p_{2;j,i-j}, \quad i \geq 0,$$

$$\pi_{n;i} = \sum_{j=0}^i p_{n;j,i-j}, \quad n \geq 3, \quad i \geq 0,$$

$$\pi_0 = p_0, \quad \pi_1 = \pi_{1;0}, \quad \pi_2 = \pi_{1;1} + \pi_{2;0},$$

$$\pi_i = \pi_{1;i-1} + \pi_{2;i-2} + \sum_{j=3}^i \pi_{j;i-j}, \quad i \geq 3.$$

Algorithm 1 Algorithm for calculation of stationary joint probability distribution

```

Initialize  $\lambda, \mu$ ;
for  $n \geq 0$  do
    Calculate  $p_n$  from Eq. (1), (2), (3);
end for
Calculate  $p_{2;0}$  from Eq. (6);
for  $n \geq 3$  do
    Calculate  $p_{n;0}$  from Eq. (4);
end for
for  $i \geq 1$  do
    Calculate  $p_{2;i}$  from Eq. (7);
    for  $n \geq 3$  do
        Calculate  $p_{n;i}$  from Eq. (5);
    end for
end for
Calculate  $p_{1;0}$  from Eq. (8);
Calculate  $p_{2;0,0}$  from Eq. (13);
for  $n \geq 3$  do
    Calculate  $p_{n;0,0}$  from Eq. (10);
end for
for  $i \geq 1$  do
    Calculate  $p_{2;i,0}$  from Eq. (15);
    for  $n \geq 3$  do
        Calculate  $p_{n;i,0}$  from Eq. (12);
    end for
end for
for  $i \geq 2$  do
    Calculate  $p_{1;i}$  from Eq. (9);
    Calculate  $p_{2;0,i}$  from Eq. (14);
    for  $n \geq 3$  do
        Calculate  $p_{n;0,i}$  from Eq. (11);
    end for
    for  $j \geq 1$  do
        Calculate  $p_{2;j,i}$  from Eq. (15);
        for  $m \geq 3$  do
            Calculate  $p_{m;j,i}$  from Eq. (12);
        end for
    end for
end for

```

In the next section we will dwell on the derivation of probability generating functions of the joint stationary distribution.

GENERATING FUNCTIONS

Though the calculation of probabilities $p_{n;i,j}$, $n \geq 2$, $i \geq 0$, $j \geq 0$ and $p_{n;i}$, $n \geq 1$, $i \geq 0$ is just a matter of computational effort due to obtained above algorithm, performance characteristics (e.g. moments and/or correlation of queue lengths in RB) are not so straightforward to obtain. Below we show that in the considered

case one can obtain expressions for probability generating functions (PGF) that ease the computation of various performance characteristics. Let us introduce the following PGF:

$$p_n(z) = \sum_{i=0}^{\infty} z^i p_{n;i}, \quad 0 \leq z \leq 1, \quad n \geq 1,$$

$$p_n(z_1, z_2) = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} z_1^{i_1} z_2^{i_2} p_{n;i_1,i_2}, \quad 0 \leq z_1 \leq 1, \\ 0 \leq z_2 \leq 1, \quad n \geq 2,$$

$$P(u, z) = \sum_{n=3}^{\infty} u^{n-3} p_n(z), \quad 0 \leq u \leq 1,$$

$$P(u, z_1, z_2) = \sum_{n=3}^{\infty} u^{n-3} p_n(z_1, z_2), \quad 0 \leq u \leq 1.$$

If one puts $z_1 = z_2 = z$ in $P(u, z_1, z_2)$, then function $P(u, z, z)$ is the double PGF of the total number of customers in buffer and servers and total number of customers in RB when all three servers are busy. By analogy $p_n(z, z)$, $n \geq 2$, is the PGF of the total number of customers total number of customers in RB and probability of total n customers in servers and buffer. In the following we will make use of PGF $P(u) = \sum_{n=3}^{\infty} u^{n-3} p_n$, $|u| \leq 1$, which, with respect to (1)–(3), equals $P(u) = p_3/(1 - \tilde{\rho}u)$.

Now, starting from equation (4), we will successively obtain relations for PGF defined above. Multiplying (4) and (5) by z^i and summing up over all i from 0 to infinity, we have:

$$(\lambda + 3\mu)p_n(z) = \lambda p_{n-1}(z) + \mu z p_{n+1}(z) + 2\mu p_{n+1}, \quad n \geq 3. \quad (16)$$

If one multiplies the previous equations by u^{n-3} and sums them up over all $n \geq 3$, then after collecting the common term, one obtains the relation for $P(u, z)$:

$$P(u, z) = \frac{\mu z p_3(z) - \lambda p_2(z)u - 2\mu[P(u) - p_3]}{\lambda u^2 - (\lambda + 3\mu)u + \mu z}. \quad (17)$$

Consider equations (6) and (7). Multiplication by z^i and summation over all possible values of i , gives the relation for $p_2(z)$:

$$(\lambda + 2\mu)p_2(z) = \lambda p_1 + 2\mu p_3 + \mu z p_3(z). \quad (18)$$

The same manipulation with equations (8) and (9). leads to relation for $p_1(z)$:

$$(\lambda + \mu)p_1(z) = \lambda p_0 + \mu p_2(z) + \mu z p_2(z, z). \quad (19)$$

Multiplying (10)–(12) by $z_1^{i_1} z_2^{i_2}$ and summing equations over all possible values of i_1 and i_2 , one gets relation for $p_n(z_1, z_2)$:

$$(\lambda + 3\mu)p_n(z_1, z_2) = \lambda p_{n-1}(z_1, z_2) + \mu p_{n+1}(z_2) + \mu z_2 p_{n+1}(z_2, z_2) + \mu z_1 p_{n+1}(z_1, z_2), \quad n \geq 3.$$

In order to obtain equation for $P(u, z_1, z_2)$ one must multiply the previous relation by u^{n-3} and again sum up over $n \geq 3$. It holds that

$$P(u, z_1, z_2) = [\lambda u^2 - (\lambda + 3\mu)u + \mu z_1]^{-1} \times \\ [\mu p_3(z_2) + \mu z_2 p_3(z_2, z_2) + \mu z_1 p_3(z_1, z_2) - \\ \mu z_2 P(u, z_2, z_2) - \mu P(u, z_2) - \lambda p_2(z_1, z_2)u]. \quad (20)$$

Finally, from (13)–(15) repeating the traditional manipulations one obtains relation for the last unknown PGF $p_2(z_1, z_2)$:

$$(\lambda + 2\mu)p_2(z_1, z_2) = \lambda p_1(z_2) + \mu p_3(z_2) + \\ \mu z_2 p_3(z_2, z_2) + \mu z_1 p_3(z_1, z_2). \quad (21)$$

In order to obtain expressions for introduced PGF one has to solve system of equations (16)–(21). It can be done as follows.

By putting $z_1 = z_2 = z$ in (20) and (21) one yields the following two equations:

$$P(u, z, z) = [\lambda u^2 - (\lambda + 3\mu)u + 2\mu z]^{-1} \times \\ [\mu p_3(z) + 2\mu z p_3(z, z) - \mu P(u, z) - \lambda p_2(z, z)u], \quad (22)$$

$$(\lambda + 2\mu)p_2(z, z) = \lambda p_1(z) + \mu p_3(z) + 2\mu z p_3(z, z). \quad (23)$$

Consider function $f_m(u, z)$, $m = 1, 2$, given by expression

$$f_m(u, z) = \lambda u^2 - (\lambda + 3\mu)u + m\mu z, \quad m = 1, 2. \quad (24)$$

Denote by $u_m = u_m(z)$ minimal and by $\hat{u}_m = \hat{u}_m(z)$ — maximal solution of the equation $f_m(u, z) = 0$, $m = 1, 2$, i.e.

$$u_m = \frac{\lambda + 3\mu - \sqrt{(\lambda + 3\mu)^2 - 4m\lambda\mu z}}{2\lambda}, \\ \hat{u}_m = \frac{\lambda + 3\mu}{\lambda} - u_m, \quad m = 1, 2.$$

Note that $0 < u_m < 1$ for $0 \leq z \leq 1$ and $m = 1, 2$. Relations (17) and (22) can be rewritten with respect to (24) in the form

$$P(u, z) = \frac{\mu z p_3(z) - \lambda u p_2(z) - 2\mu[P(u) - p_3]}{f_1(u, z)}, \quad (25)$$

$$P(u, z, z) = [f_2(u, z)]^{-1} \times \\ [2\mu z p_3(z, z) - \lambda u p_2(z, z) - \mu[P(u, z) - p_3(z)]]. \quad (26)$$

Denominator in (25) and (26) is zero at points $(u_1, z) = (u_1(z), z)$ and $(u_2, z) = (u_2(z), z)$. Since PGF $P(u, z, z)$ is analytic function in the domain $0 \leq z \leq 1$ then numerator must be zero at these points too. This leads to the following equations

$$\mu z p_3(z) - \lambda u_1 p_2(z) - 2\mu[P(u_1) - p_3] = 0, \quad (27)$$

$$2\mu z p_3(z, z) - \lambda u_2 p_2(z, z) - \mu[P(u_2, z) - p_3(z)] = 0. \quad (28)$$

Firstly we find PGF $P(u, z)$. Solution of equations (18) and (27)

$$(\lambda + 2\mu)p_2(z) - \mu z p_3(z) = \lambda p_1 + 2\mu p_3,$$

$$\mu z p_3(z) - \lambda u_1 p_2(z) = 2\mu[P(u_1) - p_3]$$

gives the expression for PGF $p_2(z)$ and $p_3(z)$ in the form:

$$p_2(z) = \frac{\lambda p_1 + 2\mu P(u_1)}{\lambda - \lambda u_1 + 2\mu},$$

$$p_3(z) = \frac{1}{\mu z} (\lambda u_1 p_2(z) + 2\mu[P(u_1) - p_3]).$$

Substitution the form of $p_2(z)$ and $p_3(z)$ in (25) and collecting the common terms, leads to expression for PGF $P(u, z)$:

$$P(u, z) = \frac{1}{(\hat{u}_1 - u)(1 - \tilde{\rho}u_1)} \times \\ \left(\frac{[(\lambda + 2\mu)p_{2,0} - \lambda p_1 \tilde{\rho}u_1]}{(\lambda + 2\mu - \lambda u_1)} + \frac{2\mu p_4}{\lambda(1 - \tilde{\rho}u)} \right).$$

Let us find the expression for $P(u, z, z)$. Solving system of equations (19), (23) and (28), one obtains the following expression for PGFs $p_1(z)$, $p_2(z, z)$ and $p_3(z, z)$:

$$p_2(z, z) = \left(\lambda - \lambda u_2 + 2\mu - \frac{\lambda \mu z}{\lambda + \mu} \right)^{-1} \times \\ \left(\frac{\lambda}{\lambda + \mu} [\lambda p_0 + \mu p_2(z)] + \mu P(u_2, z) \right),$$

$$p_1(z) = \frac{1}{\lambda + \mu} [\lambda p_0 + \mu p_2(z) + \mu z p_2(z, z)],$$

$$p_3(z, z) = \frac{1}{2\mu z} (\lambda u_2 p_2(z, z) + \mu [P(u_2, z) - p_3(z)]).$$

If one substitutes expressions for $p_1(z)$, $p_2(z, z)$ and $p_3(z, z)$ into (26) then, after collecting the common terms, one finds $P(u, z, z)$:

$$P(u, z, z) = \frac{p_2(z, z)}{(\hat{u}_2 - u)} + \frac{\mu}{\lambda(\hat{u}_2 - u)(\hat{u}_1 - u_2)} \times \\ \left(P(u, z) + \frac{2\mu p_4 \tilde{\rho}}{\lambda(1 - \tilde{\rho}u_2)(1 - \tilde{\rho}u)(1 - \tilde{\rho}u_1)} \right).$$

The last PGF to find is $P(u, z_1, z_2)$. Denominator in (20) is zero at point $(u_1(z_1), z_1)$. Since PGF $P(u, z_1, z_2)$ is analytic function in the domain $0 \leq z_1 \leq 1, 0 \leq z_2 \leq 1$ then numerator must vanish at this point. Hence it holds

$$\mu z_1 p_3(z_1, z_2) - \lambda u_1(z_1) p_2(z_1, z_2) = \\ \mu [P(u_1(z_1), z_2) - p_3(z_2)] + \\ \mu z_2 [P(u_1(z_1), z_2, z_2) - p_3(z_2, z_2)]. \quad (29)$$

From relation (21) it follows that

$$\mu z_1 p_3(z_1, z_2) = (\lambda + 2\mu) p_2(z_1, z_2) - \lambda p_1(z_2) - \\ \mu p_3(z_2) - \mu z_2 p_3(z_2, z_2).$$

Substitution of $\mu z_1 p_3(z_1, z_2)$ into (29), leads to the expression for $p_2(z_1, z_2)$:

$$p_2(z_1, z_2) = [\lambda + 2\mu - \lambda u_1(z_1)]^{-1} \times [\lambda p_1(z_2) + \mu P(u_1(z_1), z_2) + \mu z_2 P(u_1(z_1), z_2, z_2)].$$

Thus we have obtained all the unknown quantities in PGF $P(u, z_1, z_2)$ and it is determined completely. Its final expression is too cumbersome and thus omitted.

In the next section we proceed to numerical examples that depict the behaviour of queues in buffer and RB.

NUMERICAL EXAMPLE

There are several quantities related to the number of customers in the system that may be of interest. They are mean and variance of the number of customers in queue #1 and queue #2, correlation between queue size in buffer and queue #1, between queue size in buffer and queue #2 and between queue #1 and queue #2. These quantities are depicted in Fig. 2 – Fig. 4. In all examples service rate $\mu = 1$.

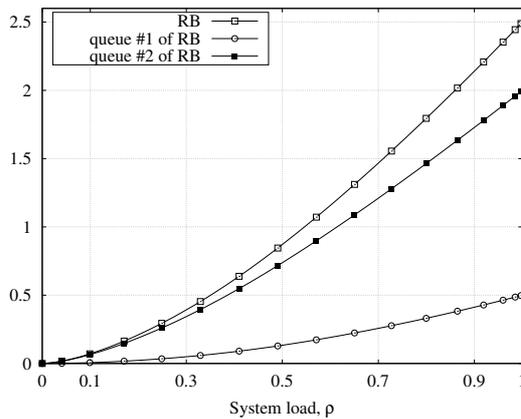


Figure 2: Mean number of customers

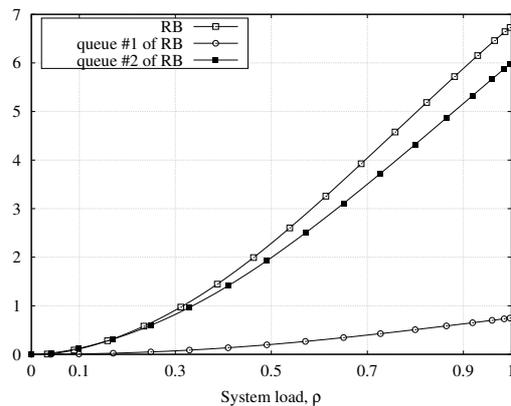


Figure 3: Variance of the number of customers

Interesting to notice from Fig. 4 that correlation between queue sizes is insignificant especially between queues of RB. This raises the question for further study about the presence of correlation between queues in RB

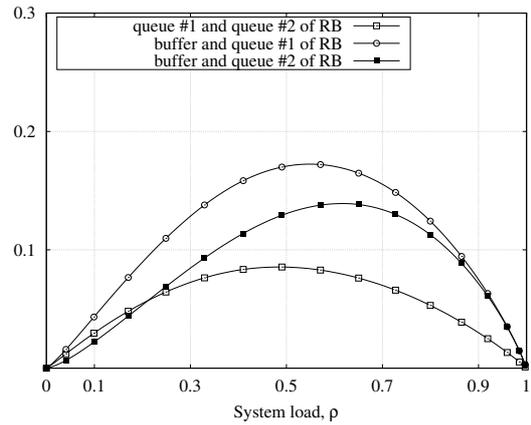


Figure 4: Coefficients of correlation

for larger values of n . The variance of the total number of customers in RB (see Fig. 3) is almost equal to the sum of variances for queue #1 and queue #2 of RB.

SUMMARY

In this study, consideration is given to $M/M/3/\infty$ queueing system with resequencing buffer of infinite capacity. Noticing that customer in reordering buffer may form two separate queues, focus is given to the study of their size distribution. Results of the thorough analysis of joint stationary distribution (both explicit and in terms of generating functions) are presented. It is shown numerically that for the all possible range of load values correlation between any queues that are formed in the system is almost insignificant. Further study will be devoted to the analysis of joint stationary distribution of queues in reordering buffer in more complex systems with possibly arbitrary number of servers.

Notes and Comments. This work was partially supported in part by the Russian Foundation for Basic Research (grants 14-07-00041, 13-07-00223).

REFERENCES

Agrawal, S., Ramaswamy, R. 1987. Analysis of the resequencing delay for $M/M/m$ systems. Proceedings of the ACM SIGMETRICS conference on Measurement and modeling of computer systems. Pp. 27–35.

Baccelli, F., Gelenbe, E., Plateau, B. 1981. An end-to-end approach to the sequencing problem. Rapports de Recherche, INRIA.

Boxma, O., Koole, G., Liu, Z. 1994. Queueing-theoretic solution methods for models of parallel and distributed systems. Performance Evaluation of Parallel and Distributed Systems Solution Methods. CWI Tract 105 and 106. Pp. 1–24.

Chowdhury, S. 1991. Distribution of the total delay of packets in virtual circuits. Proceedings of the Tenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 91). Vol. 2. Pp. 911–918.

- Chakravarthy, S., Chukova, S., Dimitrov, B. Analysis of MAP/M/2/K queueing model with infinite resequencing buffer. *Journal of Performance Evaluation*. Vol. 31. Issue 3-4. Pp. 211–228.
- Ding, Y., Sholl, H., Lipsky L. 1991. Approximations of the mean resequencing waiting time in M/GI/c systems. *Proceedings of the 11th International Conference on Distributed Computing Systems*. Pp. 28–35.
- Dimitrov, B. 1997. Queues with resequencing. A survey and recent results. *Proceedings 2-nd World Congress on Nonlinear Analysis, Theory, Methods, Applications*. Vol. 30, No. 8. Pp. 5447–5456.
- Dimitrov, B., Green Jr., D., Rykov, V., Stanchev, P. 2002. On performance evaluation and optimization problems in queues with resequencing. *Advances in Stochastic Modelling*, ed. J. Artalejo, A. Krishnamoorthy. Notable publications. Pp. 55–72.
- De Nicola, C., Pechinkin, A., Razumchik, R., 2013. Stationary Characteristics of Homogenous Geo/Geo/2 Queue with Resequencing in Discrete Time. *Proceedings of the 27th European Conference on Modelling and Simulation*. Pp. 594–600.
- Gao, Y., Zhao, Y. Q. 2012. Large Deviations for Re-Sequencing Buffer Size. *IEEE Transactions on Information Theory*. Vol. 58. No. 2. Pp. 1003–1009.
- Gogate, R., Panwar, S. 1999. Assigning customers to two parallel servers with resequencing. *IEEE Communications Letters*. Vol. 3. No. 4. Pp. 119–121.
- Huisman, T., Boucherie, R.J. (2002) The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times. *Journal of Applied Probability*. Vol. 39. No. 3. Pp. 590–603.
- Huisman, T., Boucherie, R.J. (2001) Running times on railway sections with heterogeneous train traffic. *Transportation Research Part B: Methodological*. Vol. 35. No. 3.
- Jun Li, Yifeng Zhou, Lamont, L., Minyi Huang, Zhao, Y.Q. 2010. Probabilistic Analysis of Resequencing Queue Length in Multipath Packet Data Networks. *IEEE Global Telecommunications Conference (GLOBECOM 2010)*. Pp. 1–5.
- Jain, M., Sharma, G.C. 2011. No passing multiserver queue with additional heterogeneous servers and inter-dependent rates. *5-th Canadian Conference in Applied Statistics. 20-th conference of the Forum for Interdisciplinary Mathematics -Interdisciplinary Mathematical Statistical Techniques*. Concordia University, Montreal, Quebec, Canada.
- Lelarge, M. 2008. Packet reordering in networks with heavy-tailed delays. *Mathematical Methods on Operation Research*. Vol. 67. Pp.341-371.
- Leung, K., Li, V.O.K. 2010. A resequencing model for high-speed packet-switching networks. *Journal Computer Communications*. Vol. 33. Issue 4. Pp. 443–453.
- Lelarge, M. 2008. Packet reordering in networks with heavy-tailed delays. *Mathematical Methods on Operation Research*. Vol. 67. Pp.341-371.
- Li, J., Zhao, Y.Q. 2010. Modeling and Analysis of Resequencing Delay in Selective Repeat Automatic Repeat Request. *Journal of networks*. Vol. 5. No. 7. Pp. 792–799.
- Matyushenko, S. I. 2010. Stationary characteristics of the two-channel queueing system with reordering customers and distributions of phase type. *Informatics and its Applications*. Vol. 4. Issue 4. Pp. 67–71. (in Russian).
- Min Choi, Jong Hyuk Park, Young-Sik Jeong. 2012. Revisiting reorder buffer architecture for next generation high performance computing. *The Journal of Supercomputing*. Pp. 1–12.
- Rubin, I., Zhang, Z. 1991. Message delay and queue-size analysis for circuit-switched TDMA systems. *IEEE Transactions on Communications*. Vol. 39. Issue 6. Pp. 905–914.
- Takine, T., Ren, J., Hasegawa, T. 1994. Analysis of the Resequencing Buffer in a Homogeneous M/M/2 Queue. *Performance Evaluation*. Vol. 19. Issue 4. Pp. 353–366.
- Wen-Fen, L. 2011. An Analysis of Resequencing Queue Size at Receiver on Multi-Path Transfers. *Proceedings of the International Conference on Internet Technology and Applications (iTAP)*. Pp. 1–4.
- Ye Xia, Tse, D.N.C. 2008. On the large deviations of resequencing queue size: 2-M/M/1 Case. *IEEE Transactions on information theory*. Vol. 54. No. 9. Pp. 4107–4118.
- Zheng, K., Jiao, X., Liu, M., Li, Z. 2010. An Analysis of Resequencing Delay of Reliable Transmission Protocols over Multipath. *Proceedings of the IEEE International Conference on Communications (ICC)*. Pp. 1–5.

AUTHOR BIOGRAPHIES

ILARIA CARACCIO received her degree in Mathematics in 2012 at University of Salerno. At the moment she is a Phd in Mathematics at University of Salerno. Her current research activity focuses on queuing theory. Her email address is icaraccio@unisa.it.

ALEXANDER V. PECHINKIN is a Doctor of Sciences in Physics and Mathematics and principal scientist at the Institute of Informatics Problems of the Russian Academy of Sciences, and a professor at the Peoples' Friendship University of Russia. He is the author of more than 150 papers in the field of applied probability theory. His email address is apechinkin@ipiran.ru.

ROSTISLAV V. RAZUMCHIK received his Ph.D. in Physics and Mathematics in 2011. Since then, he has worked as a senior researcher at the Institute of Informatics Problems of the Russian Academy of Sciences. His current research activities focus on queuing theory and adaptive strategies. His email address is rrazumchik@ieee.org

GENERATION OF PROBABILITY MEASURES WITH THE GIVEN SPECIFICATION OF THE SMALLEST BANS

Alexander A. Grusho, Nick A. Grusho and Elena E. Timonina
Institute of Informatics Problems of RAS
Vavilova 44-2,
119333, Moscow, Russia
Email: grusho@yandex.ru

KEYWORDS

Consistent sequences of criteria, bans of probability measures in the discrete spaces, generation of probability measures

ABSTRACT

A ban means a sequence which has zero probability in a finite space. Generation of probability models is carried out, as a rule, from simpler models by introduction of additional restrictions. However fulfilment of required properties for stochastic processes requires the proof in case of introduction of additional restrictions. In particular, the proof is required that restrictions on admissibility of trajectories don't destroy the property of being a random process that is to satisfy to the Kolmogorov's theorem.

The paper deals with conditions under which introduction of restrictions on trajectories of random sequences according to the given specification of the smallest bans again generates random process.

When the probability measure Q is generated by restrictions defined by bans we consider testing of sequence of hypotheses $H_{0,n} : Q_n$ against $H_{1,n} : P_n$, where Q possesses the specification of the smallest bans, P is a uniform measure and Q_n, P_n are projections of measures Q and P . The existence of consistency of sequence test defined by bans is investigated.

INTRODUCTION

There is a problem of anomalies search in behavior of researched processes in monitoring and control systems and in case of simulation of complex systems. One of the main tools in the solution of such problems is mathematical statistics. However in case of a nonzero error of decision-making based on observation, anomaly detection process generates a large number of false alarms (Axelson, 1999) which complicate or do impossible the analysis of the reasons of anomalies. For random processes with discrete time and finite set of states we found a case when probability of false alarm is equal zero when hypotheses are tested. But at the same time the probability of the correct decision on existence of anomalies tends to 1. This approach is based on bans of probability measures (Grusho and Timonina, 2011; Grusho et al., 2013a,b).

In the previous studies (Grusho and Timonina, 2011; Grusho et al., 2013a,b) we introduced a definition of

a ban for a probability measure on a finite space. A ban means a sequence which has zero probability in a finite space. We have shown that the notion of bans is convenient because it allows to determine the critical sets of statistical tests in the simplest way for calculation (Grusho and Timonina, 2011; Grusho et al., 2013a,b).

The analysis of anomalies in observed processes requires statistical modelling with use of an assessment of results by statistical techniques. However the application of statistical techniques is correct only in the conditions of correctly constructed probability models. In particular, if during experiment a remoted future behavior of the system is extrapolated, the probability model shall be represented by stochastic process at least (Prokhorov and Rozanov, 1993). In other words, extrapolation of any limited sections of a trajectory shall be consistent with probability distributions of process continuation. Such conditions are determined by Kolmogorov's theorem about the unique continuation of a probability measure due to the consistent finite-dimensional probability distributions. Therefore in case of study of the aberrant behavior of process the model shall meet conditions of stochastic process and restrictions on admissible trajectories of such process. Example of necessary extrapolation of a statistical output is property of consistency. At testing statistical hypotheses the property of consistency of the decision guarantees the reliability of outputs about model in case of increase in a segment of observation.

Generation of probability models is carried out, as a rule, from simpler models by introduction of additional restrictions. However fulfilment of required properties for stochastic processes requires the proof in case of introduction of additional restrictions. In particular, the proof is required that restrictions on admissibility of trajectories doesn't destroy the property of being a random process that is to satisfy to the Kolmogorov's theorem.

In this paper conditions under which introduction of restrictions on trajectories of random sequences according to the given specification of the smallest bans again generates random process are researched.

The article is structured as follows. Section 2 introduces definitions and previous results. Section 3 defines conditions for the case when probability measure is generated according to given specification of smallest bans. In Section 4 we give an example of usage of proved conditions. In Section 5 we shortly analyze applications to consistent sequences of tests.

MATHEMATICAL MODEL. BASIC DEFINITIONS AND PREVIOUS RESULTS

Let $X = \{x_1, \dots, x_m\}$ be a finite set, X^n be a Cartesian product of X , X^∞ be a set of all sequences when i -th element belongs to X . Define \mathcal{A} be a σ -algebra on X^∞ , generated by cylindrical sets. \mathcal{A} is also Borel σ -algebra in Tychonof product X^∞ , where X has a discrete topology (Bourbaki, 1968; Prokhorov and Rozanov, 1993).

On (X^∞, \mathcal{A}) a probability measure P is defined. Assume P_n is a projection of P on the first n coordinates of sequences from X^∞ . It is clear that for every $B_n \subseteq X^n$

$$P_n(B_n) = P(B_n \times X^\infty).$$

Let D_n be a support of measure P_n :

$$D_n = \{\bar{x}_n \in X^n, P_n(\bar{x}_n) > 0\}.$$

Denote $\Delta_n = D_n \times X^\infty$. The sequence $\Delta_n, n=1,2,\dots$, is nonincreasing and

$$\Delta_P = \lim_{n \rightarrow \infty} \Delta_n = \bigcap_{n=1}^{\infty} \Delta_n.$$

The set Δ_P is closed and it is a support of P .

If $\bar{x}_k \in X^k$, then \tilde{x}_{k-1} is obtained from \bar{x}_k by dropping the last coordinate.

Definition 1. Ban in measure P_n is a vector $\bar{x}_k \in X^k, k \leq n$, such that

$$P_n(\bar{x}_k \times X^{n-k}) = 0.$$

If

$$P_{k-1}(\tilde{x}_{k-1}) > 0,$$

then \bar{x}_k is the smallest ban.

If \bar{x}_k is a ban in P_n then for every $k \leq s \leq n$ and for every \bar{x}_k sequence starting with \bar{x}_k we have

$$P_s(\bar{x}_s) = 0.$$

In fact, if $P_k(\bar{x}_k) = 0$ then

$$P(\bar{x}_k \times X^\infty) = 0,$$

and

$$P(\bar{x}_k \times X^{s-k} \times X^\infty) = 0.$$

It follows that

$$P_s(\bar{x}_s) = P(\bar{x}_s \times X^\infty) \leq P(\bar{x}_k \times X^{s-k} \times X^\infty) = 0.$$

If there exists $\bar{x}_n \in X^n$ such that $P_n(\bar{x}_n) = 0$ then there exists the smallest ban.

Let for all n the support of measure P_n equals to X^n . Then the support of measure P equals to X^∞ . Let's assume that the specification of the smallest bans is a set

$$\nu = \{\nu_n, n = 1, 2, \dots\},$$

where ν_i be a number of the smallest bans of length n . The problem consists in using a measure P and the specification ν to construct a probability measure Q on space (X^∞, \mathcal{A}) at which the set of the smallest bans possesses the specification ν . For creation Q at first we

will construct the consistent system of probability measures $Q_n, n = 1, 2, \dots$, which under the Kolmogorov's theorem will unambiguously determine measure Q . Let's denote $D_n, n = 1, 2, \dots, D_n \subseteq X^n$, supports of measures Q_n , and through d_n powers of these supports. In paper (Grusho et al., 2013a) it is proved that numbers $d_n, n = 1, 2, \dots$, are unambiguously connected to the specification of ν in the following ratios

$$\nu_1 m^{n-1} + \dots + \nu_{n-1} m + \nu_n + d_n = m^n. \tag{1}$$

for all $n = 1, 2, \dots$. Thus, it is necessary to construct the consistent family of probability measures $\{Q_n\}$ which powers of supports are unambiguously determined by ratios (1).

GENERATION OF PROBABILITY MEASURES WITH THE GIVEN SPECIFICATION OF THE SMALLEST BANS

Let $D_n, n = 1, 2, \dots, D_n \subseteq X^n$, be some family of the sets fitting (1), \bar{x}_n be a arbitrary element of X^n . For all $n, n = 1, 2, \dots$, we will define functions

$$g_{n+1} : D_{n+1} \rightarrow X^n$$

as follows. For all $\bar{x}_{n+1} \in D_{n+1}, \bar{x}_{n+1} = \bar{x}_n x$,

$$g_{n+1}(\bar{x}_{n+1}) = \bar{x}_n.$$

In addition to (1) on sets $\{D_n, D_n \subseteq X^n, n = 1, 2, \dots\}$ we will superimpose the following two restrictions connected to functions g_n . For all $n, n = 1, 2, \dots$, and all $\bar{x}_{n+1} \in D_{n+1}$

$$g_{n+1}(\bar{x}_{n+1}) \in D_n, \tag{2}$$

$$g_{n+1} : D_{n+1} \xrightarrow{on} D_n. \tag{3}$$

Functions h_n are determined by analogy to functions g_n for sequence of sets X^n so that for all $n, n = 1, 2, \dots$, and all $\bar{x}_n x \in X^{n+1}$

$$h_{n+1}(\bar{x}_n x) = \bar{x}_n.$$

As measures P_n have supports X^n , it is obvious that conditions (1), (2) and (3) are executed for sequence $\{X^n, h_n\}$.

Let's take arbitrary sequence of surjective functions

$$f_n : X^n \rightarrow D_n, n = 1, 2, \dots$$

Each such function generates on X^n a probability measure Q_n with the support D_n .

Then for all n functions g_{n+1} and measures Q_{n+1} generate on X^n probability measures Q'_n with supports D_n (because functions f_{n+1} and g_{n+1} are surjective).

Theorem 1. Let arbitrary family of probability measures $\{Q_n\}$ with supports $\{D_n\}$ and family of functions $\{g_n\}$, which satisfy the conditions (2) and (3), be set. Family of probability measures $\{Q_n\}$ is the consistent if and only if for all n equalities $Q_n = Q'_n$ are executed.

Proof. We will prove the sufficiency. For consistency of measures it is enough that for all $\bar{x}_n \in X^n$

$$Q_n(\bar{x}_n) = Q_{n+1}(\bar{x}_n, X).$$

From the finiteness of probability schemes

$$Q_{n+1}(\bar{x}_n, X) = \sum_{x \in X} Q_{n+1}(\bar{x}_n x).$$

By definition

$$\begin{aligned} Q'_n(\bar{x}_n) &= Q_{n+1}(g_{n+1}^{-1}(\bar{x}_n)) = \sum_{x \in D_{n+1}} Q_{n+1}(\bar{x}_n x) = \\ &= \sum_{x \in X} Q_{n+1}(\bar{x}_n x) = Q_{n+1}(\bar{x}_n, X). \end{aligned}$$

Under the theorem condition

$$Q'_n(\bar{x}_n) = Q_n(\bar{x}_n).$$

for all $\bar{x}_n \in X^n$. From this it follows that for all $\bar{x}_n \in X^n$

$$Q_n(\bar{x}_n) = Q_{n+1}(\bar{x}_n, X).$$

Sufficiency is proved.

We will prove necessity. If $\{Q_n\}$ is the consistent family of probability measures, then for all $\bar{x}_n \in X^n$

$$Q_{n+1}(\bar{x}_n, X) = Q_n(\bar{x}_n).$$

Besides, for all $\bar{x}_n \in X^n$

$$Q'_n(\bar{x}_n) = Q_{n+1}(g_{n+1}^{-1}(\bar{x}_n)) = Q_{n+1}(\bar{x}_n, X) = Q_n(\bar{x}_n).$$

The theorem is proved.

Family of functions $\{f_n\}$ and a probability measure P generate family of probability measures $\{Q_n\}$ with supports $\{D_n\}$. Let functions $\{g_n\}$ satisfy conditions (2) and (3). Then fairly following proposition.

Corollary 1. Family of functions $\{f_n\}$ and probability measure P generate the only probability measure Q if and only if when for all $n = 1, 2, \dots$, the equality $Q_n = Q'_n$ is executed.

Theorem 2. For consistency of a set of probability measures $\{Q_n\}$ generated by functions $\{f_n\}$ and projections of a measure P on X^∞ it is enough that for all n the following diagrams are commutative

$$\begin{array}{ccc} X^n & \xleftarrow{h_{n+1}} & X^{n+1} \\ f_n \downarrow & & \downarrow f_{n+1} \\ D_n & \xleftarrow{g_{n+1}} & D_{n+1} \end{array} \quad (4)$$

where $\{g_n\}$ satisfy (2) and (3)

Proof. Each function f_n and measure P_n on X^n generate on D_n a probability distribution Q_n . Owing to consistency of projections of a measure P each function h_{n+1} generates a measure P_n from a measure P_{n+1} . Therefore it is possible to consider that the measure Q_n is generated from a measure P_{n+1} by means of composition of functions $(f_n \star f_{n+1})$.

In turn function f_{n+1} and the measure P_{n+1} generate distribution of probabilities Q_{n+1} on D_{n+1} . This measure and function g_{n+1} generate a measure Q'_n on D_n . That is the measure Q'_n on D_n is generated from measure P_{n+1} by means of a function composition $(g_{n+1} \star f_{n+1})$. Under the condition (4) functions $(f_n \star h_{n+1})$ and $(g_{n+1} \star f_{n+1})$ are equal. Therefore, these functions and measure P_{n+1} generate on D_n the same measure. That is $Q_n = Q'_n$. From here and theorems 1 consistency of family of probability measures $\{Q_n\}$ follows. The theorem is proved.

EXAMPLE OF GENERATING OF PROBABILITY MEASURES WITH THE GIVEN SPECIFICATION OF THE SMALLEST BAN

Let's consider $\nu = \{\nu_n = 1, n = 1, 2, \dots\}$, $X = \{0, 1, \dots, m-1\}$, $m > 2$, and P be a uniform measure on X^∞ . We will give an example of creation of measure Q with the specification of the smallest bans ν .

All vectors $\bar{x}_n \in X^n$ can be considered as numbers in m -dimensional numeration system. Then in each set $B_n \subseteq X^n$ there is the smallest vector \bar{x}_n considered as a number. We will build a required measure inductively. In D_1 the smallest ban we will equal to 0. Function f_1 maps X on $X \setminus \{0\}$. We will assume that D_n and f_n are defined. We will define D_{n+1} . Let \bar{x}_n^0 be the smallest number in D_n . In a set $D_n \times X$ we will define the smallest ban equals to $(\bar{x}_n^0, 0)$. Let's suppose

$$D_{n+1} = (D_n \times X) \setminus \{(\bar{x}_n^0, 0)\}.$$

Let's construct a surjective function

$$f_{n+1} : X^{n+1} \longrightarrow^{on} D_{n+1}.$$

For all $\bar{x}_n x \in X^{n+1}$ except that in which $f_n(\bar{x}_n) = \bar{x}_n^0$ and $x = 0$, we will suppose that

$$f_{n+1}(\bar{x}_n x) = (f_n(\bar{x}_n), x) \in D_n \times X.$$

Let's denote $\bar{y}_n^{(i)}$, $i = 1, \dots, k$, all members of set $f_n^{-1}(\bar{x}_n^0, 0)$. We will define

$$f_{n+1}(\bar{y}_n^{(i)}, 0) = (\bar{x}_n^0, 1) \in D_n \times X.$$

We will notice that $(\bar{x}_n^0, 1)$ is the smallest number in D_{n+1} . By definition f_n is mapping X^n on D_n . Therefore f_{n+1} maps X^{n+1} on $D_{n+1} = (D_n \times X) \setminus \{(\bar{x}_n^0, 0)\}$. Under construction D_{n+1} from D_n , for D_{n+1} there is only one smallest ban $(\bar{x}_n^0, 0)$, that is $\nu_{n+1} = 1$.

We will prove that diagrams (4) are commutative. Under construction f_n and h_{n+1}

$$f_n(\bar{y}_n^{(i)}) = \bar{x}_n^0.$$

Therefore for all $x \in X$

$$(f_n \star h_{n+1})(\bar{y}_n^{(i)}, x) = \bar{x}_n^0.$$

In case of $\bar{x}_n \neq \bar{y}_n^{(i)}$, $i = 1, \dots, k$,

$$(f_n \star h_{n+1})(\bar{x}_n, x) = f_n(\bar{x}_n).$$

Further

$$f_{n+1}(\bar{y}_n^{(i)}, 0) = (\bar{x}_n^0, 1) \in D_{n+1}, i = 1, \dots, k,$$

$$(f_n * g_{n+1})(\bar{y}_n^{(i)}, 0) = f_n(\bar{y}_n^{(i)}) = \bar{x}_n^0.$$

For elements $(\bar{y}_n^{(i)}, x), x \neq 0, i = 1, \dots, k,$

$$f_{n+1}(\bar{y}_n^{(i)}, x) = (f_n(\bar{y}_n^{(i)}, x) = (\bar{x}_n^0, x) \in D_{n+1}.$$

Therefore in case of $x \neq 0$

$$(f_{n+1} * g_{n+1})(\bar{y}_n^{(i)}, x) = \bar{x}_n^o.$$

In case of $\bar{x}_n \neq \bar{y}_n^{(i)}, i = 1, \dots, k,$ by determination f_{n+1} we have that for all $x \in X$

$$f_{n+1}(\bar{x}_n, x) = (f_n(\bar{x}_n), x) \in D_{n+1}.$$

Then under construction

$$(g_{n+1} * f_{n+1}) = (f_n * h_{n+1}).$$

Commutativity of diagrams (4) is proved.

From here it follows the existence of a measure Q on (X^∞, \mathcal{A}) with the specification of the smallest bans $\nu = \{\nu_i = 1, i = 1, 2, \dots\}.$

APPLICATION TO THE CONSISTENCY ANALYSIS

From ratios (1) we receive ratios (5)

$$d_{n+1} - md_n + \nu_{n+1} = 0, n = 1, 2, \dots \quad (5)$$

Let P be a uniform measure on (X^∞, \mathcal{A}) . Then the relation $\frac{d_n}{m^n}$ is probability of the set D_n in a measure P_n . For the specification $\nu = \{\nu_i = 1, i = 1, 2, \dots\},$ from (5) we receive the following ratios

$$\frac{d_n}{m^n} = 1 - \frac{1}{m-1} + \frac{1}{(m-1)m^n}.$$

In case of $n \rightarrow \infty$ a limit of this probability is equal to probability P of support Δ_Q of measure Q

$$P(\Delta_Q) = 1 - \frac{1}{m-1} > 0.$$

From necessary and sufficient conditions (Grusho and Timonina, 2011) of existence of consistent sequences of tests, determined by bans, for testing hypotheses $H_{0,n} : Q_n$ against $H_{1,n} : P_n$ it follows that such sequences of tests don't exist.

The vector $\bar{x}_k \in X^k$ is called the minimum ban if for all vectors $\bar{x}_n \in X^n, P_n(\bar{x}_n) > 0,$ the vector $(\bar{x}_n, \bar{x}_k) \in X^{n+k}$ is the smallest ban.

Let P be a uniform measure which generates measure Q . We will notice that if in a measure Q all bans are defined by nonempty finite set of the minimum bans, there is the consistent sequence of tests determined by bans. It follows from the fact that in measure P_n the probability of absence of the fixed vector of length k tends to 0 in case of $n \rightarrow \infty$. It means that $P(\Delta_Q) = 0$ and it is possible to apply the theorem in (Grusho and Timonina, 2011).

From this it follows that having divided both parts of ratio (1) on $m^n,$ there exists a limit in case of $n \rightarrow \infty$. We will receive the following equality

$$\frac{\nu_1}{m} + \frac{\nu_2}{m^2} + \dots + \frac{\nu_n}{m^n} \dots = 1.$$

Thus, the set of numbers $\{\frac{\nu_n}{m^n}, n = 1, 2, \dots\},$ form a distribution of probabilities on a set of natural numbers. In particular,

$$\frac{\nu_n}{m^n} \rightarrow 0, n \rightarrow \infty.$$

Ratios (5) allow receiving sufficient conditions when the specification ν of the smallest bans guarantees existence of consistent sequence of tests, determined by bans, in testing of sequence of hypotheses $H_{0,n} : Q_n$ against $H_{1,n} : P_n,$ where Q possesses the specification $\nu,$ and P is a uniform measure. From (5) equality follows

$$\frac{d_{n+1}}{m^{n+1}} = \frac{d_n}{m^n} - \frac{\nu_{n+1}}{m^{n+1}}.$$

We will suppose

$$\nu_{n+1} = \varepsilon_n md_m, 0 < \varepsilon_n < 1.$$

Then

$$\frac{d_{n+1}}{m^{n+1}} = \frac{d_n}{m^n} (1 - \varepsilon_n). \quad (6)$$

From (6) it follows

$$\frac{d_{n+1}}{m^{n+1}} = \frac{d_1}{m} \exp\left\{\sum_{i=1}^n \ln(1 - \varepsilon_n)\right\}. \quad (7)$$

In case of $n \rightarrow \infty$ the right part of (7) tends to 0 for $\varepsilon_n,$ satisfying to inequality

$$\frac{1}{n^\alpha} < \varepsilon_n < 1, 0 < \alpha \leq 1.$$

In these cases there exists (Grusho and Timonina, 2011; Grusho et al., 2013a) a consistent sequence of tests determined by bans.

CONCLUSION

In the paper the conditions of correct construction of probability models with the given specification of the smallest bans are received. Correct construction of stochastic models allows to use well developed tools of the theory of random sequences and processes in the analysis of statistical data.

We will mark that in the case of $d_n \equiv 1,$ there is no open set in the support of measure Q . In the example of section 4 in case of $\nu = \{\nu_i = 1, i = 1, 2, \dots\}$ in Δ_Q such open set exists. All open set in a uniform measure has probability bigger than 0.

Authors couldn't prove or refute a hypothesis that $P(\Delta_Q) > 0$ if and only if in case when in Δ_Q there is an open set. Correctness of this proposition would help to simplify the proof of existence of consistent sequence of tests, determined by bans, for the alternatives, which dominate a uniform measure.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grant 13-01-00480).

REFERENCES

- Axelsson, S. 1999. "The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection". In *Proc. of the 6th Conference on Computer and Communications Security*.
- Bourbaki, N. 1968. *Topologie G'enerale. Russian translation*. Science, Moscow.
- Grusho, A., N. Grusho and E. Timonina. 2013. "Consistent sequences of tests defined by bans". *Springer Proceedings in Mathematics and Statistics, Optimization Theory, Decision Making, and Operation Research Applications*, 281-291.
- Grusho, A., N. Grusho and E. Timonina. 2013. "Criteria on statistically defined bans". In *Proceedings of 27th European Conference on Modelling and Simulation* (May 27-30, 2013, Alesund, Norway). Digitaldruck Pirrot GmbH, Dudweiler, Germany, 610-614.
- Grusho, A. and E. Timonina. 2011. "Prohibitions in discrete probabilistic statistical problems". *Discrete Mathematics and Applications* 21, No.3, 275-281.
- Prokhorov, U.V.; and U.A. Rozanov. 1993. *Theory of probabilities*. Science, Moscow.

AUTHOR BIOGRAPHIES

ALEXANDER A. GRUSHO was born 28.08.1946, Doctor of sciences (Math), Professor of Moscow State University, leading scientist of Institute of Informatics Problems of Russian Academy of Sciences. His email is grusho@yandex.ru.

NICK A. GRUSHO was born 15.06.1982, PhD (Math), senior scientist of Institute of Informatics Problems of Russian Academy of Sciences. His email is info@itake.ru.

ELENA E. TIMONINA was born 27.02.1952, Doctor of sciences (Tech), Professor, leading scientist of Institute of Informatics Problems of Russian Academy of Sciences. Her email is eltimon@yandex.ru.

ON THE DEVELOPMENT OF AN INFORMATION TECHNOLOGY FOR PLASMA TURBULENCE RESEARCH

Andrey Gorshenin

Institute of Informatics Problems, Russian Academy of Sciences
Vavilova str., 44/2, Moscow, Russia
MIREA, Faculty of Information Technology
Email: agorshenin@ipiran.ru

Victor Korolev

Lomonosov Moscow State University
Leninskie Gory, Moscow, Russia
Institute of Informatics Problems, Russian Academy of Sciences
Email: victoryukorolev@yandex.ru

Dmitry Malakhov

Nina Skvortsova
Prokhorov General Physics Institute,
Russian Academy of Sciences
Vavilova str., 38, Moscow, Russia
Email: mukudori@mail.ru

Sergey Shorgin

Institute of Informatics Problems, Russian Academy of Sciences
Vavilova str., 44/2, Moscow, Russia
Email: sshorgin@ipiran.ru

Victor Kuzmin

"Wi2Geo LLC", Russia
Email: shadesilent@gmail.com

KEYWORDS

Information technology, Turbulent plasma, Simulation, Probabilistic models, Spectrum.

ABSTRACT

The paper deals with an example of real information technology developed for the examination of specific structures of plasma turbulence by the spectral analysis. The mathematical basis is a probabilistic approach using a special simulation algorithm to construct sample for the probabilistic modeling. To describe the fine structure of stochastic processes, finite mixtures of various probability distributions are used.

In the paper, the general structure of the developed information technology including mathematical models, algorithms and software realization is considered. Procedures for finding statistical estimators of the unknown parameters of model are based on modifications of EM algorithm. Examples of application of the developed technology in an important area of modern physics are presented.

INTRODUCTION

To create effective and safety sizeable scientific plants using concept of tokamaks (e.g., International Thermonuclear Experimental Reactor, ITER) adequate models of plasma functioning have to be developed. And the first step for it is the construction of mathematical models, algorithms and software for special laboratory plants.

To estimate process parameters, spectrum measured by spectrometer, spectrograph or analog-to-digital converter must be split into components (continuous spectrum, bands, single components). Spectral analysis is one of the traditional tools of a signal processing in the plasma turbulence. But the decomposition of a plasma turbulence spectrum is an ill-posed problem. The unique solution exists only under additional assumptions about the object's structure (Lochte-Holtgreven, 1968; Akhmanov, 2004). So, it is impossible to obtain important spectral information about the functioning of plasma turbulence by the traditional approach implying spectrum's approximation by Kolmogorov-Obukhov model or shot (fluctuation) noise model.

However, the spectra must be analyzed since it makes possible to reveal the set of important physical parameters: the type of instability, the mechanism of turbulence formation, the proportion between plasma fluctuations and plasma structures, e.g., ion-acoustic solitons and drift vortices.

To overcome difficulties, special probabilistic procedure for plasma spectrum decomposition was implemented in the paper (Gorshenin et al., 2011). The spectrum is interpreted as density of unknown distribution. Then a test sample with pre-specified size is simulated using special bootstrap-type procedure. To this sample we apply an approach based on compound Cox process model and thus an approximation to the spectrum can be obtained. The choice of Cox processes for turbulence modelling is based on known empirical and theoretical results (see, e.g., the book (Korolev and Skvortsova, 2006)). Surely, one of the most important problems is that of the choice and optimization of computational methods for the

estimation of unknown parameters.

Under some experimental conditions, the efficiency of the methodology was demonstrated. The spectra were successfully decomposed into the components. Moreover, new results confirmed earlier models. For example, only few components are significant for plasma turbulence functioning despite of huge overall number of processes in plasma.

The paper represents the information technology created for automating of experimental data processing using the ideas mentioned above. In the further sections we discuss the models, algorithms and improvement techniques included in the information technology. Experimental samples for various plasma conditions from L-2M stellarator (Pshenichnikov et al., 2005) were used as data sets for probabilistic modelling.

MATHEMATICAL MODELS

The design of methods for the analysis of stochastic processes is very important for the evaluation of turbulence characteristics when conditions of the plasma confinement are changed. In this section we discuss mathematical approach for probabilistic modeling in the implemented information technology.

The spectrum is interpreted as probability density function of unknown probability distribution \mathcal{F} . We can obtain any α -quantile (values of α are chosen pseudo-random from uniform distribution on the interval $[0, 1]$) by solving the equation

$$\mathcal{F}(x_\alpha) = \alpha.$$

A sample \mathcal{X} from unknown distribution \mathcal{F} could be simulated for any predetermined size. Investigating low-frequency fluctuations of the plasma it was found that the processes in plasma turbulence could be adequately described by compound Cox processes (Korolev, 2011). So using this model we can assume that $\mathcal{F}(x)$ is a finite mixture of probability distributions (e.g., normal, gamma) with unknown parameters, i.e.:

$$\mathcal{F}(x) = \sum_{i=1}^k p_i F(x, a_i, b_i), \quad (1)$$

$$\sum_{i=1}^k p_i = 1, p_i \geq 0, \quad (2)$$

where $F(\cdot)$ denotes some type of cumulative distribution functions; $k \geq 1$ is a known natural number; $a_i, b_i, i = 1, \dots, k$, are parameters of distribution under correct conditions (e.g., standard deviations for normal mixtures are strictly positive). Quantities $F(x, a_i, b_i), i = 1, \dots, k$, are components of mixture; p_1, \dots, p_k are weights; k is a number of components

in mixture. Values $p_i, a_i, b_i, i = 1, \dots, k$, are usually unknown and should be estimated by sample. Resulting mixture approximates the spectrum.

For the normal mixtures $\mathcal{F}(x)$ in equality (1) has the following form:

$$\mathcal{F}(x) = \sum_{i=1}^k p_i \Phi\left(\frac{x - a_i}{\sigma_i}\right),$$

where

$$a_i \in \mathbf{R}, \sigma_i > 0, i = 1, \dots, k,$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{t^2}{2}\right\} dt, \quad x \in \mathbf{R},$$

and conditions (2) hold.

For the gamma mixtures $\mathcal{F}(x)$ in equality (1) has the following form:

$$\mathcal{F}(x) = \sum_{i=1}^k p_i F_{r_i, \theta_i}(x),$$

where

$$r_i > 0, \theta_i > 0, i = 1, \dots, k,$$

$$F_{r, \theta}(x) = \int_0^x \frac{\theta^r t^{r-1} e^{-\theta t}}{\Gamma(r)} dt, \quad x > 0,$$

and conditions (2) hold.

To find unknown parameters, various modifications of EM algorithm (Dempster et al., 2011) are used in the information technology. For the case of normal mixtures, the estimators can be found in (Korolev, 2011), gamma-mixtures estimators were described, e.g., in (Gorshenin and Korolev, 2013). Characteristics of grid modifications are considered in (Gorshenin et al., 2013). Combination of different methods and special statistical procedures make it possible to realize an acceptable computation accuracy and to obtain a reasonable time for different types of spectra.

The initial data are the one- and two-sided Fourier spectra. The size of the simulated samples is 100000 elements. Maximum number of components for EM algorithm equals 6. As a stopping criterion we use the relation

$$\max \left| \theta^{(m)} - \theta^{(m-1)} \right| < \varepsilon,$$

where $\theta^{(m)}$ is a vector of estimated parameters at m -th iteration step, and the ε is an accuracy. The computing accuracy ε equals 10^{-8} (multiple experiments with the EM algorithm demonstrate the fact that this value provides both the correct results and a reasonable time). Decomposition of the spectrum into components gives opportunity to understand the behavior of different types of fluctuations and structures in the plasma.

ALGORITHMS AND SOFTWARE

In this section we discuss issues of functioning and interactions of software modules representing the computational part of the information technology.

In modern researches with huge sizes of experimental data and critical required processing time it is impossible to imagine an analysis without a creation of specialized information technologies. To decompose plasma spectra into the components the corresponding solution based on the mathematical model of compound Cox processes was suggested.

Having developed mathematical tools, new results for adequacy and velocity were obtained. It predetermined set of software's algorithms. With built-in MATLAB fourth-generation programming language the computational module for spectrum's analysis was created and optimized for plasma physics research.

The information technology consists of mathematical, computational and visualization tools and the user interface for the researchers usability. Let us consider the structure of the developed software (see Fig. 1).

Experimental data are external files with access via software interface. The program can be run with a help both the standard MATLAB console (there is the main processing function) and special interface to specify parameters of methods. The first way is suitable for "developer mode", while the most of users definitely prefer interface which hides the details of implementation and simplifies working with a program.

There are three logical modules in the software's functional:

1. the simulation module (we simulate a sample for modelling);
2. the estimation module (we estimate unknown parameters by the sample using different methods and obtain approximating mixture);
3. the visualization module (we plot various figures).

Obviously, each of these steps must be performed consecutively (in Fig. 1 it is demonstrated by the dashed arrows), but you can start at any point (in Fig. 1 possible execution paths are shown by solid arrows from the module "Interface"). It was realized to process data previously saved on the disk without re-simulation, re-estimating of parameters or both of them.

The test sample is formed as described in the previous section. To estimate unknown parameters the following methods can be used by user's choice in the estimation module:

1. EM algorithm for normal mixtures;

2. EM algorithm for gamma mixtures;
3. grid method for normal mixtures;
4. grid method for gamma mixtures.

The methods for the normal mixtures can be used for one- and two-sided spectra, the methods for the gamma mixtures should be applied in case of one-sided spectra.

EM algorithm is a quite universal method for finding maximum likelihood estimates. It allows to achieve a balance between velocity and quality of approximation. However, if we have some additional information about parameters, the grid methods can be more efficient. So, both types of methods were included in the final software.

The problem of correct determination of number of components arises during an approximation by finite distributions mixtures. Therefore, most powerful tests (see, e.g., (Gorshenin, 2011)) and values of the Lo statistic (Lo et al., 2001) should be used for these purposes. The program checks significance of components with low weights by different criteria and returns the number of components in the mixture. The significance level of criterion is a parameter of the method.

The visualization module forms graphical output. According to parameter estimates the approximating curve and the constituent components are plotted to provide an intuitive interpretation of the results. The graphs allow to determine a number of components (i.e. processes in turbulent plasma) and specify some physical characteristics.

Moreover, analyzed experimental data are the series of spectra obtained under various conditions. To demonstrate evolution of spectra in time you can plot three-dimensional surface in the visualization module. A profile in concrete time moment represents approximating mixture for the corresponding spectrum. Three-dimensional graphics can be rotated and zoomed by mouse. The images could be saved in the graphic (JPEG, PNG, EPS) and MATLAB's formats (FIG).

Mathematical algorithms were approved on the test samples with known characteristics. As part of the development of information technology the special procedures to improve efficiency and accuracy were created. Using different algorithms and special techniques in the united software we made a detailed analysis of data for sustainable series of spectra obtained in low-frequency plasma turbulence at the edge and in the center of the plasma filament in the L-2M stellarator.

The software can be run on computers with an installed MATLAB as well as without it (but in this situation freeware MATLAB Compiler Runtime have to be installed).

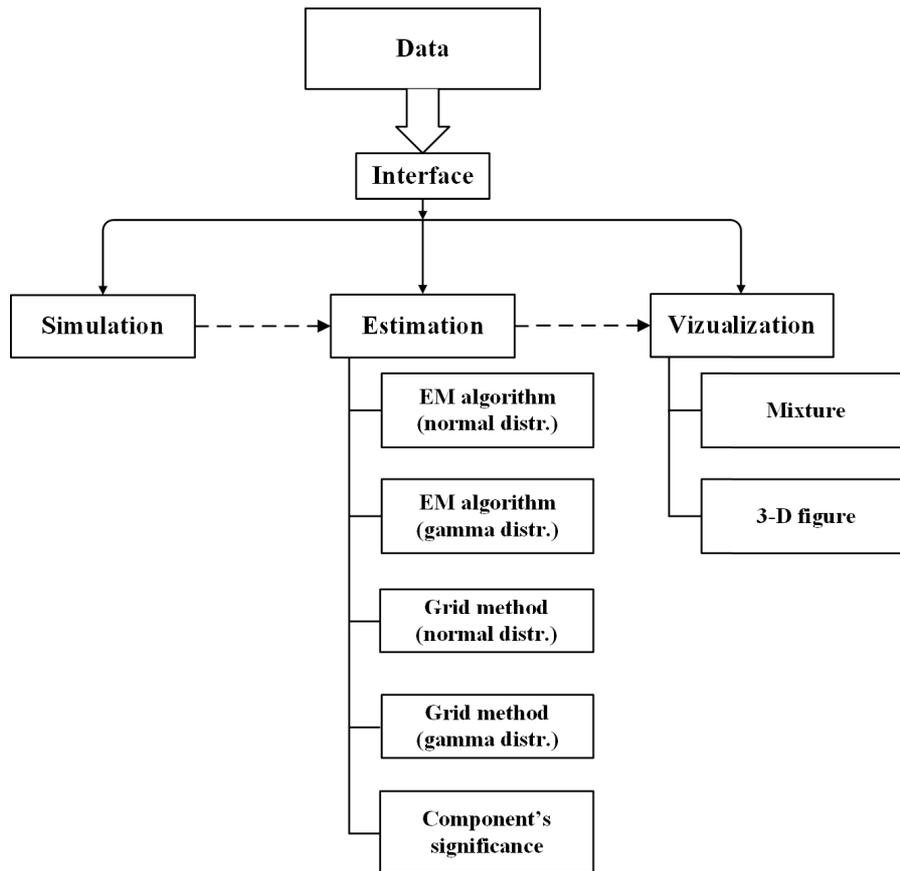


Figure 1: Structure of the information technology

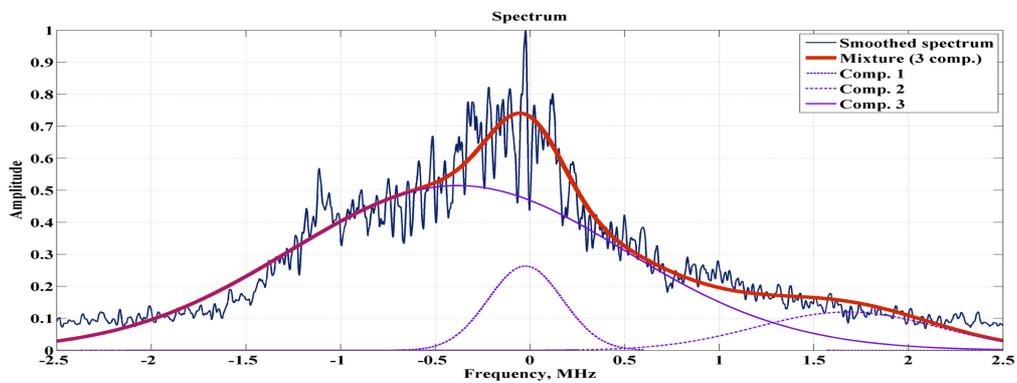


Figure 2: Two-sided spectrum

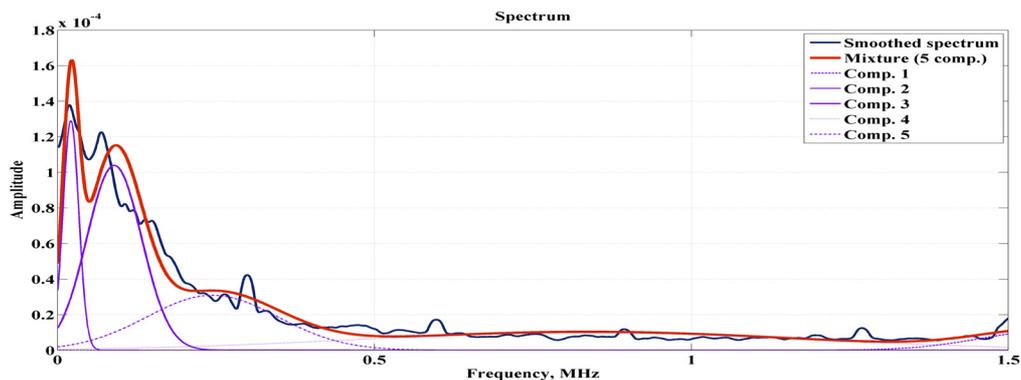


Figure 3: One-sided spectrum

APPLICATION

In this section, we demonstrate examples of applying the implemented information technology to the real data which represent measurements in the laboratory stellarator L-2M.

Figures 2-3 demonstrate examples of the software's output. We can see initial spectrum (thin solid line), approximating mixture (bold line) and the components. The x-coordinate corresponds to frequencies, the y-coordinate shows the amplitude of spectrum. There are three dominant components with different non-zero weights. It corresponds to the energy transfer between different types of turbulence. Most of the trends in the original spectrum are approximated by modelled mixture quite good.

Figure 2 demonstrates example of processing for two-sided spectrum in the certain time moment of plasma discharge in the stellarator. There are 3 forming components with non-zero weights.

Figure 3 demonstrates example of processing for one-sided spectrum in the L-2M. There are 5 forming components with non-zero weights. As seen on figures the number of forming components is finite and moreover it is a modest as compared with total number of stochastic processes in plasma turbulence. Both figures demonstrates spectrum's decomposition into the normal components.

The main present reached physical results by the applying of information technology to the real data are following:

1. the decomposition of the spectra: the form of harmonics in amplitude spectra was found out;
2. the specification of correct number of the forming processes in plasma turbulence;
3. the reveal of recurrence of stochastic processes with typical mean frequencies of spectrum's half-width;
4. the determination values of the radial electric field fluctuations, phase velocities, etc.

POSSIBLE IMPROVEMENTS

In this section we discuss possible ways for the information technology to improve a productivity. Undoubtedly, it can be embodied by special programming solutions, e.g., we can realize computational modules using any low-level programming language. The source code might be very effective and fast but too difficult for programming and especially for debugging. Moreover, it needs much time and in fact you will create a new information technology. So in this section we consider approaches of possible optimizations of the existent technology.

The first way is simply to use more actual and efficient hardware. For example, we have used the Intel

Core i7 CPU and obtained up to 3 times acceleration of working with one spectrum comparing our previous hardware. Moreover, modern CPUs have more than one logical core and so you can process multiple data sets simultaneously. The ratio between velocity and time is not linear, nevertheless you obtain significant acceleration without special techniques!

The second way logically follows the first one in terms of parallelism. Modern integrated development environments (e.g., built-in MATLAB IDE) support mechanisms for automatization of parallel computing for a source code. Using special directives, program can work faster without wide modifications of the code.

The third way is based on new hardware ideas and creating special source code for these purposes. It leads to computing on GPUs, clusters, etc. At that, GPU solutions are not so expensive as clusters and supercomputers. The fact that nowadays CPUs are created with integrated graphics chips (e.g., AMD Fusion, Intel Ivy Bridge, etc.) demonstrates extent of perspectiveness of GPU computing. The world leading GPU producers offer special solutions for researchers in different areas (CUDA technology by NVIDIA, ATI Stream Technology by AMD). It should be noted that in modern GPUs the number of cores equals from several hundred to thousands ones. Obviously, their performance may be extremely high for various complex computational problems in the areas with the critical requirements for accuracy and processing time. Surely, one of the most important problems is a creation of an effective software that would be able to use the full power of the hardware solutions. In fact the optimal application performance on multi-core systems can be achieved through rational use of program threads for the correct allocation of subproblems. Threads execution can be optimized for running on a different physical cores.

CONCLUSION

As mentioned above the proposed methodology is quite successful for problems of approximation of a spectrum and its decomposition into the components. The number of components and the forms are invariable in specific regions in the series. The resulting number of significant components in the decomposition of the spectrum adjusts with the physical essence of the processes under study because the amount of significant processes in real data is a modest.

The use of non-parametric estimation procedures in plasma physics is very prospective for the analysis of short-wavelength plasma turbulence. With the help of a physical interpretation of the component it is possible to create more precise models of the turbulent plasma functioning.

Thus, the created information technology presents an effective software tool of plasma processes research. Moreover, just now we obtained informal outcomes for the real data sets.

Due to the informational properties of probability distributions, some known physical aspects the problem of modelling of spectra by mixtures of another distributions can be very interesting and perspective. Although it seems to be an object of special independent research.

The problem of spectrum components decomposition is a relevant in other physical spheres, e.g., quantum decoding of the harmonic spectra of diatomic molecules like a titanium (II) oxide TiO (Hermann et al., 2001).

The implemented methods contributes to the application of mathematical methods and computational techniques in the physics of plasma turbulence. The created information technology can be used as software basis of processing experimental data in the data centers. To use the solution in real plants like ITER some software and hardware improvements including mentioned in previous section are needed. But success methodology for laboratory plant allows to expect potentially interesting results in this case too.

ACKNOWLEDGEMENTS

The research is supported by the Grant of the President of the Russian Federation (project MK-4103.2014.9) and by the Russian Foundation for Basic Research (projects 12-07-00115a, 12-07-00109a, 14-07-00041a).

REFERENCES

- Akhmanov, S.A.; and S.Yu. Nikitin. 2004. *Physical Optics*. Nauka, Moscow.
- Dempster, A., N. Laird and D. Rubin. 1977. "Maximum likelihood estimation from incompleting data." *Journal of the Royal Statistical Society. Series B* 39, No.1, 1-38.
- Gorshenin, A.K. 2011. "Testing of statistical hypotheses in the splitting component model." *Moscow University Computational Mathematics and Cybernetics* 35, No.4, 176-183.
- Gorshenin, A.K., V.Yu. Korolev, D.V. Malakhov and N.N. Skvortsova. 2011. "Analysis of fine stochastic structure of chaotic processes by kernel estimators." *Matematicheskoe modelirovanie* 23, No.4, 83-89.
- Gorshenin, A., V. Korolev. 2013. "Modeling of statistical fluctuations of information flows by mixtures of gamma distributions". In *Proceedings of 27th European Conference on Modelling and Simulation*

(May 27-30, 2013, Alesund, Norway). Digitaldruck Pirrot GmbH, Dudweiler, Germany, 569-572.

- Gorshenin, A., V. Korolev, V. Kuzmin and A. Zeifman. 2013. "Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution". In *Proceedings of 27th European Conference on Modelling and Simulation* (May 27-30, 2013, Alesund, Norway). Digitaldruck Pirrot GmbH, Dudweiler, Germany, 565-568.
- Hermann, J., A. Perrone and C. Dutouquet. 2001. "Analyses of the TiO - γ system for temperature measurements in a laser-induced plasma." *J. Phys. B: At. Mol. Opt. Phys.* 34, 153-164.
- Korolev, V. Yu. 2011. *Probabilistic and Statistical Methods of Decomposition of Volatility of Chaotic Processes*. Moscow University Publishing House, Moscow.
- Korolev, V.Yu.; and N.N. Skvortsova (Eds). 2006. *Stochastic Models of Structural Plasma Turbulence*. VSP, Utrecht.
- Lo, Y., N.R. Mendell and D.B. Rubin. 2001. "Testing the number of components in a normal mixture." *Biometrika* 88, No.3, 767-778.
- Lochte-Holtgreven, W. 1968. *Plasma Diagnostics*. North-Holland, Amsterdam; Interscience (Wiley), New York.
- Pshenichnikov, A.A., L.V. Kolik, N.I. Malykh, A.E. Petrov et al. 2005. "The use of Doppler reflectometry in the L-2M stellarator." *Plasma Phys. Rep.* 31, No.7, 554-561.
- AUTHOR BIOGRAPHIES**
- ANDREY GORSHENIN** is Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; associate professor, Faculty of Information Technology, Moscow State Institute of Radio Engineering, Electronics and Automation. His e-mail address is: agorshenin@ipiran.ru
- VICTOR KOROLEV** is Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences. His e-mail address is: victoryukorolev@yandex.ru

DMITRY MALAKHOV is Candidate of Science (PhD) in physics and mathematics, Prokhorov General Physics Institute, Russian Academy of Sciences. His e-mail address is: 89199945791@mail.ru

NINA SKVORTSOVA is Doctor of Science in physics and mathematics, Prokhorov General Physics Institute, Russian Academy of Sciences. Her e-mail address is: mukudori@mail.ru

SERGEY SHORGIN is Doctor of Science in physics and mathematics, professor, deputy director, Institute of Informatics Problems, Russian Academy of Sciences. His e-mail address is: sshorgin@ipiran.ru

VICTOR KUZMIN is Head of Development, "Wi2Geo LLC", Russia. His e-mail address is: shadesilent@gmail.com

ON TRUNCATIONS FOR SZK MODEL

Alexander Zeifman

Vologda State University, S.Orlova, 6, Vologda, Russia

Institute of Informatics Problems, Russian Academy of Sciences

Institute of Territories Socio-Economic Development, Russian Academy of Sciences

Yakov Satin and Galina Shilova

Vologda State University

S.Orlova, 6, Vologda, Russia

Victor Korolev and Vladimir Bening

Moscow State University, Leninskie Gory, Moscow, Russia

Institute of Informatics Problems, Russian Academy of Sciences

Sergey Shorgin

Institute of Informatics Problems, Russian Academy of Sciences

Vavilova str., 44-2, Moscow, Russia

KEYWORDS

Markovian queueing models; SZK model; inhomogeneous Markov processes; truncations

ABSTRACT

We consider a class of inhomogeneous Markovian queueing models with batch arrivals and group services. Bounds on the truncation errors in weak ergodic case are obtained. Two concrete queueing models are studied as examples.

INTRODUCTION

The problem of existence and construction of limiting characteristics for inhomogeneous continuous-time Markov chains is important for queueing applications, see for instance, (Granovsky and Zeifman 2004, Zeifman et al. 2006). Calculation of the limiting characteristics for birth-death process via truncations was firstly mentioned in (Zeifman 1991) and was considered in details in (Zeifman et al. 2006).

About two decades ago V. Kalashnikov suggested that in some cases one can obtain uniform (in time) error bounds of truncation, and in (Zeifman et al. 2014b) we prove this conjecture for inhomogeneous birth-death processes.

A new class of Markovian time-inhomogeneous queueing models with batch arrivals and group services was introduced and studied in our recent papers (Satin et al. 2013, Zeifman et al. 2014a). Bounds of the rate of convergence, perturbation bounds, and approximations via truncations were studied in these papers.

Here we consider this model and obtain uniform in time error bounds of truncation.

Consider a time-inhomogeneous continuous-time Markovian queueing model ("SZK model") on the state

space $E = \{0, 1, \dots\}$ with possible batch arrivals and group services.

Let $X = X(t)$, $t \geq 0$ be a queue-length process for the queue, $p_{ij}(s, t) = Pr \{X(t) = j | X(s) = i\}$, $i, j \geq 0$, $0 \leq s \leq t$, be transition probabilities for $X = X(t)$, and $p_i(t) = Pr \{X(t) = i\}$ be its state probabilities. Throughout the paper we assume that

$$\Pr (X(t+h) = j | X(t) = i) = \begin{cases} q_{ij}(t)h + \alpha_{ij}(t, h), & \text{if } j \neq i, \\ 1 - \sum_{k \neq i} q_{ik}(t)h + \alpha_i(t, h), & \text{if } j = i, \end{cases} \quad (1)$$

where all $\alpha_i(t, h)$ are $o(h)$ uniformly in i , i. e. $\sup_i |\alpha_i(t, h)| = o(h)$. We also assume $q_{i, i+k}(t) = \lambda_k(t)$, $q_{i, i-k}(t) = \mu_k(t)$ for any $k > 0$. In other words, we suppose that the arrival rates $\lambda_k(t)$ and the service rates $\mu_k(t)$ do not depend on the length of queue. In addition, we assume that $\lambda_{k+1}(t) \leq \lambda_k(t)$ and $\mu_{k+1}(t) \leq \mu_k(t)$ for any k and almost all $t \geq 0$. Applying our standard approach (see details in (Granovsky and Zeifman 2004, Zeifman 1995a, Zeifman et al. 2006, 2008)) we suppose in addition, that all intensity functions are linear combinations of a finite number of locally integrable on $[0, \infty)$ nonnegative functions. Moreover we assume

$$\lambda_k(t) \leq \lambda_k, \quad \mu_k(t) \leq m_k, \quad (2)$$

for any k and almost all $t \geq 0$, and

$$L_\lambda = \sum_{i=1}^{\infty} \lambda_i < \infty, \quad L_\mu = \sum_{i=1}^{\infty} \mu_i < \infty. \quad (3)$$

Then the probabilistic dynamics of the process is represented by the forward Kolmogorov system:

$$\frac{d\mathbf{p}}{dt} = A(t)\mathbf{p}(t), \quad (4)$$

where

$$A(t) = \begin{pmatrix} a_{00}(t) & \mu_1(t) & \mu_2(t) & \cdots & \mu_r(t) & \cdots \\ \lambda_1(t) & a_{11}(t) & \mu_1(t) & \cdots & \mu_{r-1}(t) & \cdots \\ \lambda_2(t) & \lambda_1(t) & a_{22}(t) & \cdots & \mu_{r-2}(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_r(t) & \lambda_{r-1}(t) & \lambda_{r-2}(t) & \cdots & a_{rr}(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \quad (5)$$

and $a_{ii}(t) = -\sum_{k=1}^i \mu_k(t) - \sum_{k=1}^{\infty} \lambda_k(t)$.

Throughout the paper by $\|\cdot\|$ we denote the l_1 -norm, i. e. $\|\mathbf{x}\| = \sum |x_i|$, and $\|B\| = \sup_j \sum_i |b_{ij}|$ for $B = (b_{ij})_{i,j=0}^{\infty}$.

Let Ω be a set all stochastic vectors, i. e. l_1 vectors with nonnegative coordinates and unit norm. Hence we have $\|A(t)\| = 2 \sum_{k=1}^{\infty} (\lambda_k(t) + \mu_k(t)) \leq 2(L_\lambda + L_\mu)$ for almost all $t \geq 0$. Hence the operator function $A(t)$ from l_1 into itself is bounded for almost all $t \geq 0$ and locally integrable on $[0; \infty)$. Therefore we can consider (4) as a differential equation in the space l_1 with bounded operator.

It is well known, see (Daleckij and Krein 1974), that the Cauchy problem for differential equation (4) has a unique solutions for an arbitrary initial condition, and $\mathbf{p}(s) \in \Omega$ implies $\mathbf{p}(t) \in \Omega$ for $t \geq s \geq 0$.

Denote by $E(t, k) = E\{X(t) | X(0) = k\}$ the mean (the mathematical expectation) of the process at the moment t under the initial condition $X(0) = k$.

Recall that a Markov chain $X(t)$ is called *weakly ergodic*, if $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \rightarrow 0$ as $t \rightarrow \infty$ for any initial conditions $\mathbf{p}^*(0), \mathbf{p}^{**}(0)$, where $\mathbf{p}^*(t)$ and $\mathbf{p}^{**}(t)$ are the corresponding solutions of (4).

A Markov chain $X(t)$ has the *limiting mean* $\varphi(t)$, if $\lim_{t \rightarrow \infty} (\varphi(t) - E(t, k)) = 0$ for any k .

Let $\{d_i\}$, $i = 1, 2, \dots$ be an increasing sequence of positive numbers, $d_1 = 1$. Put

$$W = \inf_{i \geq 1} \frac{d_i}{i}, \quad g_i = \sum_{n=1}^i d_n, \quad (6)$$

Denote

$$\alpha_i(t) = -a_{ii}(t) - \sum_{k \geq 1} \lambda_k(t) \frac{d_{k+i}}{d_i} - \sum_{k=1}^{i-1} (\mu_{i-k}(t) - \mu_i(t)) \frac{d_k}{d_i}, \quad (7)$$

and

$$\alpha(t) = \inf_{i \geq 1} \alpha_i(t). \quad (8)$$

Assume that for some positive K

$$d_1 \lambda_1 + (d_1 + d_2) \lambda_2 + \cdots \leq K, \quad (9)$$

By introducing $p_0(t) = 1 - \sum_{i \geq 1} p_i(t)$, from (4) we obtain the equation

$$\frac{d\mathbf{z}}{dt} = B(t)\mathbf{z}(t) + \mathbf{f}(t), \quad (10)$$

where $\mathbf{f}(t) = (\lambda_1, \lambda_2, \dots)^\top$, $\mathbf{z}(t) = (p_1, p_2, \dots)^\top$,

$$B = (b_{ij}(t))_{i,j=1}^{\infty} = \begin{pmatrix} a_{11} - \lambda_1 & \mu_1 - \lambda_1 & \cdots & \mu_{r-1} - \lambda_1 & \cdots \\ \lambda_1 - \lambda_2 & a_{22} - \lambda_2 & \cdots & \mu_{r-2} - \lambda_2 & \cdots \\ \lambda_2 - \lambda_3 & \lambda_1 - \lambda_3 & \cdots & \mu_{r-3} - \lambda_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{r-1} - \lambda_r & \lambda_{r-2} - \lambda_r & \cdots & a_{rr} - \lambda_r & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \quad (11)$$

see detailed discussion in (Granovsky and Zeifman 2004, Zeifman 1995a, Zeifman et al. 2006, 2008). Let D be upper triangular matrix,

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \quad (12)$$

and let l_{1D} be the corresponding space of sequences

$$l_{1D} = \{\mathbf{z} = (p_1, p_2, \dots)^\top | \|\mathbf{z}\|_{1D} \equiv \|D\mathbf{z}\|_1 < \infty\}.$$

Consider equation (10) in the space l_{1D} , where $B(t)$ and $\mathbf{f}(t)$ are locally integrable on $[0, +\infty)$. Consider the logarithmic norm of operator function $B(t)$, see the respective motivation in (Van Doorn et al. 2010, Granovsky and Zeifman 2004, Zeifman et al. 2008) and detailed proofs in (Zeifman 1995b). Recall that the logarithmic norm of operator function $B(t)$ is defined as

$$\gamma(B(t)) = \lim_{h \rightarrow +0} h^{-1} (\|I + hB(t)\| - 1).$$

The important inequality

$$\|V(t, s)\| \leq \exp \int_s^t \gamma(B(u)) du$$

holds, where $V(t, s) = V(t)V^{-1}(s)$ is the Cauchy operator of equation (10). Further, for an operator function from l_1 to itself we have the simple formula

$$\gamma(B(t)) = \sup_j \left(b_{jj}(t) + \sum_{i \neq j} |b_{ij}(t)| \right).$$

Hence we obtain the following bound for the logarithmic norm of the operator function $B(t)$:

$$\gamma(B(t))_{1D} = \gamma(DB(t)D^{-1}) = \sup_{i \geq 1} \{-\alpha_i(t)\} = -\alpha(t), \quad (13)$$

where

$$DBD^{-1} = \begin{pmatrix} a_{11} & (\mu_1 - \mu_2) \frac{d_1}{d_2} & (\mu_2 - \mu_3) \frac{d_1}{d_3} & \cdots & (\mu_{r-1} - \mu_r) \frac{d_1}{d_r} & \cdots \\ \lambda_1 \frac{d_2}{d_1} & a_{22} & (\mu_1 - \mu_3) \frac{d_2}{d_3} & \cdots & (\mu_{r-2} - \mu_r) \frac{d_2}{d_r} & \cdots \\ \lambda_2 \frac{d_3}{d_1} & \lambda_1 \frac{d_3}{d_2} & a_{33} & \cdots & (\mu_{r-3} - \mu_r) \frac{d_3}{d_r} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{r-1} \frac{d_r}{d_1} & \lambda_{r-2} \frac{d_r}{d_2} & \lambda_{r-3} \frac{d_r}{d_3} & \cdots & a_{rr} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \quad (14)$$

Therefore

$$\|V(t, s)\|_{1D} \leq e^{-\int_s^t \alpha(u) du}. \quad (15)$$

TRUNCATIONS

Consider the ‘‘truncated’’ process $X_{N-1}(t)$ on the state space $E_{N-1} = \{0, 1, \dots, N-1\}$ with the corresponding intensity matrix

$$A_{N-1} = \begin{pmatrix} b_{00} & \mu_1 & \mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{N-1} \\ \lambda_1 & b_{11} & \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{N-2} \\ \lambda_2 & \lambda_1 & b_{22} & \mu_1 & \mu_2 & \cdots & \mu_{N-3} \\ \cdots & & & & & & \\ \lambda_{N-1} & \lambda_{N-2} & \lambda_{N-3} & \cdots & \lambda_2 & \lambda_1 & b_{N-1, N-1} \end{pmatrix} \quad (16)$$

where $b_{ii}(t) = -\sum_{k=1}^i \mu_k(t) - \sum_{k=1}^{N-1-i} \lambda_{N-1-k}(t)$.

Instead of (4), for $X_{N-1}(t)$ we have the following forward Kolmogorov system:

$$\frac{d\mathbf{p}^*}{dt} = A_{N-1}(t)\mathbf{p}^*. \quad (17)$$

Similarly, instead of (10), we obtain the correspondent reduced system for the truncated process in the form:

$$\frac{d\mathbf{z}^*}{dt} = B_{N-1}(t)\mathbf{z}^*(t) + \mathbf{f}_{N-1}(t), \quad (18)$$

where $\mathbf{f}_{N-1}(t) = (\lambda_1, \dots, \lambda_{N-1}, 0, \dots)^\top$, $\mathbf{z}^*(t) = (p_1, p_2, \dots, p_{N-1})^\top$,

$$B_{N-1} = (b_{ij}^*(t))_{i,j=1}^{N-1} = \quad (19)$$

$$\begin{pmatrix} b_{11} - \lambda_1 & \mu_1 - \lambda_1 & \cdots & \mu_{N-1} - \lambda_1 \\ \lambda_1 - \lambda_2 & b_{22} - \lambda_2 & \cdots & \mu_{N-2} - \lambda_2 \\ \lambda_2 - \lambda_3 & \lambda_1 - \lambda_3 & \cdots & \mu_{N-3} - \lambda_3 \\ \cdots & & & \\ \lambda_{N-2} - \lambda_{N-1} & \lambda_{N-3} - \lambda_{N-1} & \cdots & b_{N-1, N-1} - \lambda_{N-1} \end{pmatrix}$$

Below we will identify the finite vector with entries, say, $(a_1, \dots, a_{N-1})^\top$ and the infinite vector with the same first $N-1$ coordinates and the others equal to zero.

For bounding of the truncation error we rewrite (18) in the following form:

$$\frac{d\mathbf{z}^*}{dt} = B(t)\mathbf{z}^*(t) + \mathbf{f}(t) - \hat{B}(t)\mathbf{z}^*(t) - \hat{\mathbf{f}}(t), \quad (20)$$

where $\hat{B}(t) = B(t) - B_{N-1}(t)$, and $\hat{\mathbf{f}}(t) = \mathbf{f}(t) - \mathbf{f}_{N-1}(t)$.

Then we have

$$\mathbf{z}^*(t) = V(t)\mathbf{z}^*(0) + \int_0^t V(t, \tau)\mathbf{f}(\tau) d\tau - \int_0^t V(t, \tau)\hat{B}(\tau)\mathbf{z}^*(\tau) d\tau - \int_0^t V(t, \tau)\hat{\mathbf{f}}(\tau) d\tau, \quad (21)$$

Hence, if $\mathbf{z}(0) = \mathbf{z}^*(0)$, then the sum of first and second summands gives us $\mathbf{z}(t)$, and we obtain in any norm

$$\|\mathbf{z}(t) - \mathbf{z}^*(t)\| \leq \left\| \int_0^t V(t, \tau)\hat{B}(\tau)\mathbf{z}^*(\tau) d\tau \right\| + \left\| \int_0^t V(t, \tau)\hat{\mathbf{f}}(\tau) d\tau \right\| \leq \int_0^t \|V(t, \tau)\| \|\hat{B}(\tau)\mathbf{z}^*(\tau)\| d\tau + \int_0^t \|V(t, \tau)\| \|\hat{\mathbf{f}}(\tau)\| d\tau. \quad (22)$$

We have

$$\hat{\mathbf{f}}(t) = \mathbf{f}(t) - \mathbf{f}_{N-1}(t) = (0, \dots, 0, \lambda_N(t), \lambda_{N+1}(t), \dots)^\top, \quad (23)$$

and

$$\|\hat{\mathbf{f}}(t)\|_{1D} = \|D\hat{\mathbf{f}}(t)\|_1 = (d_1 + \dots + d_N)\lambda_N(t) + (d_1 + \dots + d_{N+1})\lambda_{N+1}(t) + \dots \leq g_N\lambda_N + g_{N+1}\lambda_{N+1} + \dots \rightarrow 0 \text{ as } N \rightarrow \infty, \quad (24)$$

in accordance with assumption (9).

On the other hand,

$$\hat{B}(t)\mathbf{z}^*(t) = (B(t) - B_{N-1}(t))\mathbf{z}^*(t) = ((a_{11}(t) - b_{11}(t))p_1^*(t), \dots, (a_{N-1, N-1}(t) - b_{N-1, N-1}(t))p_{N-1}^*(t))^\top, \quad (25)$$

and

$$\begin{aligned} \|\hat{B}(t)\mathbf{z}^*(t)\|_{1D} &= \|D(B(t) - B_{N-1}(t))\mathbf{z}^*(t)\|_1 = \\ &= d_1(b_{11}(t) - a_{11}(t))p_1^*(t) + (d_1 + d_2)(b_{22}(t) - a_{22}(t))p_2^*(t) + \dots + \\ &= (d_1 + \dots + d_{N-1})(b_{N-1, N-1}(t) - a_{N-1, N-1}(t))p_{N-1}^*(t) = \\ &= d_1 \sum_{k \geq N-1} \lambda_k(t)p_1^*(t) + (d_1 + d_2) \sum_{k \geq N-2} \lambda_k(t)p_2^*(t) + \dots + \\ &= (d_1 + \dots + d_{N-1}) \sum_{k \geq 1} \lambda_k(t)p_{N-1}^*(t). \end{aligned} \quad (26)$$

Let now a sequence $\{d_i\}$ be such that $1 = d_1 \leq d_2 \leq \dots$, and the following two assumptions hold:

$$\|V(t, s)\|_{1D} \leq M e^{-a(t-s)} \quad (27)$$

for any $0 \leq s \leq t$, and some positive numbers M and a ,

and

$$\|V(t, s)\|_{1D^*} \leq M^* e^{-a^*(t-s)} \quad (28)$$

for any $0 \leq s \leq t$, some positive numbers M^* and a^* , where

$$D^* = \begin{pmatrix} d_1^2 & d_1^2 & d_1^2 & \cdots \\ 0 & d_2^2 & d_2^2 & \cdots \\ 0 & 0 & d_3^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \quad (29)$$

Let, in addition, there exist a positive number K^* such that

$$d_1^2\lambda_1 + (d_1^2 + d_2^2)\lambda_2 + \dots \leq K^*. \quad (30)$$

Now we try to estimate $\|\hat{B}(t)\mathbf{z}^*(t)\|_{1D}$.

Firstly,

$$\begin{aligned} \|\mathbf{z}^*(t)\|_{1D^*} &\leq \|V(t)\|_{1D^*} \|\mathbf{z}^*(0)\|_{1D^*} + \\ &\int_0^t \|V(t, \tau)\|_{1D^*} \|\mathbf{f}(\tau)\|_{1D^*} d\tau \leq \\ &M^* e^{-a^* t} \|\mathbf{z}^*(0)\|_{1D^*} + \frac{K^* M^*}{a^*}, \end{aligned} \quad (31)$$

because $\|\mathbf{f}(t)\|_{1D^*} \leq K^*$ for almost all $t \geq 0$.

Put $X(0) = X_{N-1}(0) = 0$, then $\mathbf{z}^*(0) = \mathbf{0}$, hence

$$\|\mathbf{z}^*(t)\|_{1D^*} \leq \frac{K^* M^*}{a^*}, \quad (32)$$

for any $t \geq 0$.

Suppose for definiteness that N is odd. All $p_i^*(t) \geq 0$, then

$$\begin{aligned} \|\mathbf{z}^*(t)\|_{1D^*} &= \sum_{i \geq 1} p_i^*(t) \sum_{k=1}^i d_k^2 \geq \\ &\sum_{i \geq \frac{N-1}{2}} d_i^2 p_i^*(t) \geq \sum_{i=\frac{N-1}{2}}^{N-1} d_i^2 p_i^*(t), \end{aligned} \quad (33)$$

On the other hand we have the following bound:

$$\begin{aligned} &d_1 \sum_{k \geq N-1} \lambda_k(t) p_1^*(t) + \\ &(d_1 + d_2) \sum_{k \geq N-2} \lambda_k(t) p_2^*(t) + \dots + \\ &(d_1 + \dots + d_{N-1}) \sum_{k \geq 1} \lambda_k(t) p_{N-1}^*(t) \leq \\ &(d_1 + \dots + d_{\frac{N-1}{2}}) \sum_{k \geq \frac{N-1}{2}} \lambda_k(t) \sum_{k=1}^{\frac{N-1}{2}} p_k^*(t) + \\ &\sum_{k \geq 1} \lambda_k(t) \left((d_1 + \dots + d_{\frac{N-1}{2}}) p_{\frac{N-1}{2}}^*(t) + \dots + \right. \\ &\quad \left. (d_1 + \dots + d_{N-1}) p_{N-1}^*(t) \right). \end{aligned} \quad (34)$$

Denote by $\Lambda_K = \sum_{k \geq K} \lambda_k$.

Then we obtain from (26), (33) and (34):

$$\begin{aligned} \|\hat{B}(t)\mathbf{z}^*(t)\|_{1D} &\leq g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} \sum_{k=1}^{\frac{N-1}{2}} p_k^*(t) + \\ &L_\lambda \left(g_{\frac{N-1}{2}} p_{\frac{N-1}{2}}^*(t) + \dots + g_{N-1} p_{N-1}^*(t) \right) \leq \\ &g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} + L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \left(d_{\frac{N-1}{2}}^2 p_{\frac{N-1}{2}}^*(t) + \right. \\ &\quad \left. \dots + d_{N-1}^2 p_{N-1}^*(t) \right) \leq g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} + \\ &L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \|\mathbf{z}^*(t)\|_{1D^*} \leq g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} + L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \frac{K^* M^*}{a^*}, \end{aligned} \quad (35)$$

for any $t \geq 0$.

Finally, we have from (22), (24), and (35) the following bound of truncation error:

$$\begin{aligned} \|\mathbf{z}(t) - \mathbf{z}^*(t)\| &\leq \int_0^t M e^{-a(t-\tau)} \left(g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} \right. \\ &\quad \left. + L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \frac{K^* M^*}{a^*} \right) d\tau + \\ &\int_0^t M e^{-a(t-\tau)} (g_N \lambda_N + g_{N+1} \lambda_{N+1} + \dots) d\tau \leq \\ &\frac{M}{a} \left(g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} + L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \frac{K^* M^*}{a^*} + \right. \\ &\quad \left. g_N \lambda_N + g_{N+1} \lambda_{N+1} + \dots \right). \end{aligned} \quad (36)$$

Now let l_{1E} be the space of sequences,

$$l_{1E} = \left\{ z = (p_1, p_2, \dots)^\top \mid \|z\|_{1E} \equiv \sum n |p_n| < \infty \right\}.$$

Then (36) and well-known inequality $\|z\|_{1E} \leq W^{-1} \|z\|_{1D}$ (see, for instance, Zeifman et al 2006), imply the following statement.

Theorem 1. Let (9), (27), (28), (30) be fulfilled. Then $X(t)$ is exponentially weakly ergodic, has the limiting mean, say, $E(t, 0)$, and the following bound of truncation error holds:

$$\begin{aligned} |E(t, 0) - E_{N-1}(t, 0)| &\leq \\ &\frac{M}{aW} \left(g_{\frac{N-1}{2}} \Lambda_{\frac{N-1}{2}} + L_\lambda \frac{g_{N-1}}{d_{\frac{N-1}{2}}^2} \frac{K^* M^*}{a^*} + \right. \\ &\quad \left. g_N \lambda_N + g_{N+1} \lambda_{N+1} + \dots \right). \end{aligned} \quad (37)$$

for any $t \geq 0$, where $E_{N-1}(t, k) = E \{ X_{N-1}(t) \mid X_{N-1}(0) = k \}$ is the mean (the mathematical expectation) of the truncated process at the moment t under initial condition $X_{N-1}(0) = k$.

EXAMPLES

1. Consider the simplest analogue of $M_t/M_t/S$ queue for a queueing system with group services, see Section 4 in (Zeifman et al. 2014a).

Namely, we suppose that the arrival intensity of a customer to the queue is $\lambda_1(t) = \lambda(t) = 1 + \sin 2\pi t$, $\lambda_i(t) = 0$ for any $i > 1$ and any $t \geq 0$. Let the intensity of departure (servicing) of a group of k customers be $\mu_k(t) = \frac{\mu(t)}{k}$ if $k \leq S < \infty$, and $\mu_i(t) = 0$ for any $i > S$ and any $t \geq 0$, where $\mu(t) = 3 + \cos 2\pi t$.

Let $X = X(t)$, $t \geq 0$ be a queue-length process for the queue. For definiteness, put $S = 10^{12}$. For this queue bound (37) looks essentially simpler:

$$|E(t, 0) - E_{N-1}(t, 0)| \leq \frac{L_\lambda M K^* M^* g_{N-1}}{a a^* W d_{\frac{N-1}{2}}^2}. \quad (38)$$

Here $L_\lambda = 2$.

Put $d = \sqrt{2}$, and $d_{k+1} = d^k$. Then $K = K^* = 2$, $W = \sqrt{2}$. Now, in (8) we have

$$\alpha(t) \geq \mu(t) - (d-1)\lambda(t), \quad \alpha^*(t) \geq \mu(t) - (d^2-1)\lambda(t), \quad (39)$$

therefore we can apply bound (15) and obtain $a \geq 2$, $a^* \geq 2$, $M \leq 2$, $M^* \leq 2$.

Finally, we have $g_{N-1} \leq 2^{\frac{N}{2}}$, $d_{\frac{N-1}{2}}^2 = 2^{N-2}$, and the following estimate for the error of truncation:

$$|E(t, 0) - E_{N-1}(t, 0)| \leq 16 \cdot 2^{-\frac{N}{2}}, \quad (40)$$

which does not depend on t in contrast to the estimate (50) in (Zeifman et al. 2014a).

Applying the approach of (Zeifman et al. 2006) we can find the approximation of the mathematical expectation of the length of queue, see the following figures.

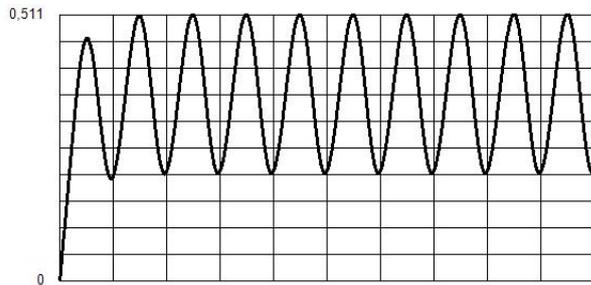


Figure 1: First example, approximation of the mean $E(t, 0)$ on $[0, 10]$ with an error 10^{-3} .

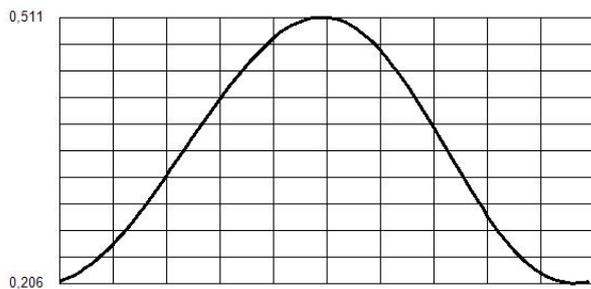


Figure 2: First example, approximation of the mean $E(t, 0)$ on $[9, 10]$ with an error 10^{-3} .

2. Consider a simplest queuing system with one server and batch arrivals, see Section 5 in (Zeifman et al. 2014a). Namely, let $\mu_1(t) = \mu(t) = 3 + \cos 2\pi t$ be the service rate of a customer, $\mu_i(t) = 0$ for any $i > 1$ and any $t \geq 0$. Let $\lambda_k(t) = \frac{\lambda(t)}{4^k}$, where $\lambda(t) = 1 + \sin 2\pi t$ be the arrival intensity of k customers to the queue. We

have $L_\lambda < 1$. Put also $d = \sqrt{2}$, and $d_{k+1} = d^k$. Then $K \leq 2$, $K^* = 2$, $W = \sqrt{2}$. Now, in (8) we have

$$\alpha(t) \geq \mu(t) - (d-1)\lambda(t), \quad \alpha^*(t) \geq \mu(t) - (d^2-1)\lambda(t), \quad (41)$$

therefore we can apply bound (15) and obtain $a \geq \frac{1}{2}$, $a^* \geq \frac{1}{2}$, $M \leq 2$, $M^* \leq 2$, $\Lambda_K \leq 4^{-K}$. We have also $g_{N-1} \leq 2^{\frac{N}{2}}$, $d_{\frac{N-1}{2}}^2 = 2^{N-2}$. For this queue we have instead of (37) the following estimate for the error of truncation:

$$|E(t, 0) - E_{N-1}(t, 0)| \leq 10^2 \cdot 2^{-\frac{N}{2}}, \quad (42)$$

which does not depend on t in contrast to the estimate (52) in (Zeifman et al. 2014a).

Now, applying the approach of (Zeifman et al. 2006) we can find the approximation of the mathematical expectation of the length of queue, see the following figures.

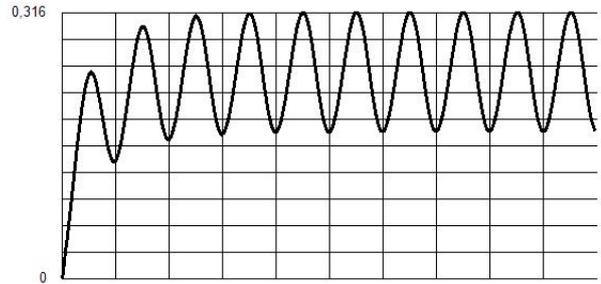


Figure 3: Second example, approximation of the mean $E(t, 0)$ on $[0, 10]$ with an error 10^{-3} .

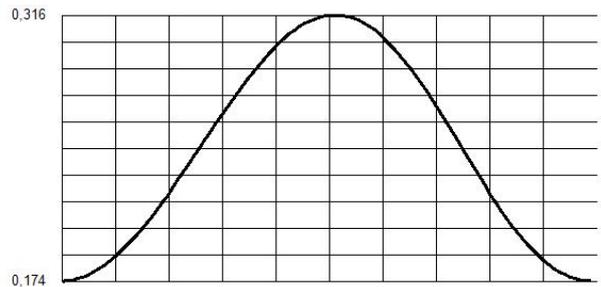


Figure 4: Second example, approximation of the mean $E(t, 0)$ on $[9, 10]$ with an error 10^{-3} .

Acknowledgement. This work was supported by the Russian Foundation for Basic Research, projects no. 14-07-00041, 13-07-00223, 12-07-00115, 12-07-00109.

REFERENCES

- Daleckij, Ju.L. and M.G. Krein. 1974. *Stability of solutions of differential equations in Banach space*. Amer. Math. Soc. Transl. 43.
- Van Doorn, E.A., A.I. Zeifman and T.L. Panfilova. 2010. "Bounds and asymptotics for the rate of convergence of birth-death processes." *Theory Probab. Appl.* 54, 97–113.
- Granovsky, B.L. and A.I. Zeifman. 2004. "Nonstationary Queues: Estimation of the Rate of Convergence." *Queueing Syst.* 46, 363–388.
- Satin, Ya. A.; A. I. Zeifman and A. V. Korotysheva. 2013. "On the rate of convergence and truncations for a class of Markovian queueing systems." *Theory. Prob. Appl.* 57, 529–539.
- Zeifman, A. I. 1991. "Qualitative properties of inhomogeneous birth and death processes." *J. Math. Sci.* 57, 3217–3224 (The Russian original paper was published in 1988).
- Zeifman, A. I. 1995a. "Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes." *Stoch. Proc. Appl.* 59, 157–173.
- Zeifman, A. I. 1995b. "On the estimation of probabilities for birth and death processes." *J. App. Probab.* 32, 623–634.
- Zeifman, A.; S. Leorato; E. Orsingher; Ya. Satin and G. Shilova. 2006. "Some universal limits for nonhomogeneous birth and death processes." *Queueing Syst.* 52, 139–151.
- Zeifman, A. I.; V.E. Bening and I.A. Sokolov. 2008. *Continuous-time Markov chains and models*. Elex-KM, Moscow (in Russian).
- Zeifman, A.; V. Korolev; A. Korotysheva, Y. Satin and V.Bening. 2014a. "Perturbation bounds and truncations for a class of Markovian queues." *Queueing Syst.* 76, 205–221.
- Zeifman, A.I.; Y. Satin; V. Korolev and S. Shorgin. 2014b. "On truncations for weakly ergodic non-stationary birth and death processes." *Int. J. Appl. Math. Comp. Sci.*

AUTHOR BIOGRAPHIES

ALEXANDER ZEIFMAN Doctor of Science in physics and mathematics; professor, Head of Department of Applied Mathematics, Vologda State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; principal scientist, Institute of Territories Socio-Economic Development, Russian Academy of Sciences. His email is a_zeifman@mail.ru and his personal webpage at <http://uni-vologda.ac.ru/zai/eng.html>.

YAKOV SATIN is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. His email is yacovi@mail.ru.

GALINA SHILOVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. Her email is shgn@mail.ru.

VICTOR KOROLEV is Doctor of Science in physics and mathematics, professor, Head of Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

VLADIMIR BENING is Doctor of Science in physics and mathematics; professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences. His email is bening@yandex.ru.

SERGEY YA. SHORGIN is Doctor of Science in physics and mathematics, professor, Deputy Director, Institute of Informatics Problems, Russian Academy of Sciences. His email is sshorgin@ipiran.ru.

ON CONVERGENCE OF THE DISTRIBUTIONS OF RANDOM SUMS AND STATISTICS CONSTRUCTED FROM SAMPLES WITH RANDOM SIZES TO EXPONENTIAL POWER LAWS

Victor Korolev

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University
Institute for Informatics Problems, Russian Academy of Sciences

M. E. Grigoryeva
Parexel International

Alexander Zeifman
Vologda State University, Russia
Institute of Informatics Problems, Russian Academy of Sciences
Institute of Territories Socio-Economic Development, Russian Academy of Sciences

KEYWORDS

Exponential power distribution; symmetric stable law; one-sided stable distribution; scale mixture of normal laws; random sum; sample with random size; mixed Poisson distribution

ABSTRACT

Limit theorems are proved establishing criteria of convergence of the distributions of random sums and statistics constructed from samples with random sizes to exponential power laws.

INTRODUCTION. EXPONENTIAL POWER DISTRIBUTIONS

Let $0 < \alpha \leq 2$. *Exponential power distribution* is the absolutely continuous distribution defined by its Lebesgue probability density

$$\ell_\alpha(x) = \frac{\alpha}{2\Gamma(\frac{1}{\alpha})} \cdot e^{-|x|^\alpha}, \quad -\infty < x < \infty. \quad (1)$$

To simplify the notation and calculation, here and in what follows we will use a single parameter α in representation (1) since this parameter is in some sense characteristic and determines the shape of distribution (1). With $\alpha = 1$ relation (1) defines the classical Laplace distribution with zero mean and variance 2. With $\alpha = 2$ relation (1) defines the normal (Gaussian) distribution with zero mean and variance $\frac{1}{2}$.

The class of distributions (1) was introduced and studied in (Subbotin 1923). Along with the term *generalized Laplace distribution* going back to the original paper (Subbotin 1923) at least four other different terms are used for distribution (1). For example, in (Box and Tiao 1973) this distribution is called *exponential power*

distribution, in (Evans et al. 2000) and (Leemis and McQueston 2008) it is called *generalized error distribution*, in (Morgan 1996) the term *generalized exponential distribution* is used whereas in (Nadaraja 2005) and (Varanasi and Aazhang 1989) this distribution is called *generalized normal* and, *generalized Gaussian* respectively. Distributions of type (1) are widely used in Bayesian analysis and various applications from astronomy to signal and image processing.

In applied probability there is a convention, apparently historically going back to the book (Gnedenko and Kolmogorov 1954), according to which a model distribution can be regarded as reasonable and/or justified enough only if it is an *asymptotic approximation*, that is, there exist a more or less simple setting and the corresponding limit theorem in which the model under consideration is a limit distribution. An interrelation of this convention with the principle of the non-decrease of uncertainty in closed systems was traced in the book (Gnedenko and Korolev 1996). A well-known reasonable numerical characteristic of uncertainty is the entropy. As we have already seen, with $0 < \alpha \leq 2$ the exponential power distribution is a scale mixture of normal laws. At the same time, the normal distribution has the maximum (differential) entropy among all laws with the finite second moment whose support is the whole real axis. According to the principle of the non-decrease of entropy which often manifests itself in probability theory in the form of limit theorems for sums of independent random variables (see (Gnedenko and Korolev 1996)), if the modeled system were information-isolated from the outer medium, then the observed statistical distributions of its characteristics would have been very close to the normal law. But since any mathematical model by its definition cannot make account of all the factors which influence the current state or the evolution of the modeled system, then the param-

eters of this normal law vary depending on the evolution of the medium exogenous with respect to the system under consideration. In other words, these parameters should be regarded as random depending on the information flows between the system and exogenous medium. Thus, in many situations reasonable mathematical models of statistical regularities of the behavior of the observed characteristics of complex systems should have the form of mixtures of normal laws, the particular case of which is the exponential power distribution (1).

Probably, by now the simplicity of representation (1) has been the main (at least, important) reason for using the exponential power distributions in many applied problems as a heavy-tailed (for $0 < \alpha < 2$) alternative to the normal law. The "asymptotic" reasons of possible adequacy of this model have not been provided yet. In this paper we will demonstrate that the exponential power distribution can be limiting in rather simple limit theorems for regular statistics constructed from samples with random sizes, in particular, in the scheme of random summation. Hence, along with the normal law, this distribution can be regarded as an asymptotic approximation for the distributions of some processes, say, similar to (non-homogeneous) random walks.

The main part of the paper is organized as follows. Normal mixture representation is studied in Section 2. We obtain a criterion of convergence of the distributions of random sums to exponential power distributions in Section 3. In Section 4 we consider a criterion of convergence of the distributions of regular statistics constructed from samples with random sizes to exponential power distributions. In Section 5 we discuss the obtained results. Finally, in Section 6 we obtain estimates on the rate of convergence of the distributions of random sums to exponential power laws.

Normal mixture representation

In (West 1987) it was noticed that for $0 < \alpha \leq 2$ the distributions of type (1) are representable as scale mixtures of normal laws (also see (Choy and Smith 1997)). For the sake of convenience of further references here we retell the proof of this result from (West 1987) in other terms.

By $G_{\alpha,\theta}(x)$ and $g_{\alpha,\theta}(x)$ we will respectively denote the distribution function and probability density of the strictly stable law with characteristic exponent α and parameter θ defined by the characteristic function

$$g_{\alpha,\theta}(t) = \exp \left\{ -|t|^\alpha \exp \left\{ -\frac{i\pi\theta\alpha}{2} \operatorname{sign} t \right\} \right\}, \quad t \in \mathbb{R}, \quad (2)$$

with $0 < \alpha \leq 2$, $|\theta| \leq \theta_\alpha = \min\{1, \frac{2}{\alpha} - 1\}$ (see, e. g., (Zolotarev 1986)). Let

$$h_{\alpha/2}(z) = \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \cdot \frac{g_{\alpha/2,1}(z)}{\sqrt{z}}, \quad z \geq 0,$$

$$w_{\alpha/2}(z) = \frac{h_{\alpha/2}(z^{-1})}{z^2} = \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \cdot \frac{g_{\alpha/2,1}(z^{-1})}{z^{3/2}}, \quad z \geq 0.$$

Below, in the proof of lemma 1, we will show that both $h_{\alpha/2}(z)$ and $w_{\alpha/2}(z)$ are probability densities. Assume that all the random variables mentioned in this paper are defined on the same probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ which is rich enough. The symbol $\stackrel{d}{=}$ denotes the coincidence of distributions. If $V_{\alpha/2}$ and $U_{\alpha/2}$ are non-negative absolutely continuous random variables with densities $h_{\alpha/2}(z)$ and $w_{\alpha/2}(z)$, respectively, then, as is easily seen,

$$U_{\alpha/2} \stackrel{d}{=} V_{\alpha/2}^{-1}. \quad (3)$$

It is well known that if $\zeta_{\alpha,\theta}$ is a random variable with the stable distribution corresponding to characteristic function (2), then $E|\zeta_{\alpha,\theta}|^p < \infty$ for every $p < \alpha$. Therefore, from the definition of the density $h_{\alpha/2}(z)$ it follows that $EV_{\alpha/2}^p < \infty$ for any $p < (\alpha + 1)/2$ and hence, (3) implies that $EU_{\alpha/2}^q < \infty$ for any $q > 0$.

The distribution functions corresponding to the densities $l_\alpha(x)$, $h_{\alpha/2}(z)$ and $w_{\alpha/2}(z)$ will be denoted by the capital letters $L_\alpha(x)$, $H_{\alpha/2}(z)$ and $W_{\alpha/2}(z)$, respectively. The standard normal distribution function ($\alpha = 2$) and its density will be respectively denoted $\Phi(x)$ and $\varphi(x)$,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^x \varphi(z) dz.$$

LEMMA 1. *If $0 < \alpha \leq 2$, then the exponential power distribution (1) is a scale mixture of normal laws:*

$$L_\alpha(x) = \int_0^\infty \Phi(x\sqrt{z}) dH_{\alpha/2}(z), \quad x \in \mathbb{R}, \quad (4)$$

$$L_\alpha(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{z}}\right) dW_{\alpha/2}(z), \quad x \in \mathbb{R}. \quad (5)$$

PROOF. From (2) it follows that the characteristic function of the symmetric ($\theta = 0$) strictly stable distribution has the form

$$g_{\alpha,0}(t) = e^{-|t|^\alpha}, \quad t \in \mathbb{R}. \quad (6)$$

On the other hand, it is well known that the symmetric strictly stable distribution with parameter α is a scale mixture of normal laws in which the one-sided ($\theta = 1$) stable law with parameter $\alpha/2$ is the mixing distribution:

$$G_{\alpha,0}(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{z}}\right) dG_{\alpha/2,1}(z), \quad x \in \mathbb{R} \quad (7)$$

(see, e. g., (Zolotarev 1986), theorem 3.3.1). Write relation (7) in terms of characteristic functions with the account of (6):

$$e^{-|t|^\alpha} = \int_0^\infty \exp\left\{-\frac{t^2 z}{2}\right\} g_{\alpha/2,1}(z) dz. \quad (8)$$

Then, re-denoting the argument $t \mapsto x$ and making some formal transformations of equality (8), we obtain

$$\ell_\alpha(x) = \frac{\alpha}{2\Gamma(\frac{1}{\alpha})} e^{-|x|^\alpha} = \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \int_0^\infty \frac{\sqrt{z}}{\sqrt{2\pi}} \exp\left\{-\frac{x^2 z}{2}\right\} = \frac{g_{\alpha/2,1}(z)}{\sqrt{z}} dz = \int_0^\infty \sqrt{z} \varphi(x\sqrt{z}) h_{\alpha/2}(z) dz. \quad (9)$$

It can be easily verified that $h_{\alpha/2}(z)$ is the probability density of a nonnegative random variable. Indeed, for any $z > 0$ we have

$$\int_{-\infty}^\infty \sqrt{z} \varphi(x\sqrt{z}) dx = 1.$$

Therefore it follows from (9) that

$$1 = \int_{-\infty}^\infty \ell_\alpha(x) dx = \int_{-\infty}^\infty \int_0^\infty \sqrt{z} \varphi(x\sqrt{z}) h_{\alpha/2}(z) dz dx = \int_0^\infty h_{\alpha/2}(z) \left(\int_{-\infty}^\infty \sqrt{z} \varphi(x\sqrt{z}) dx \right) dz = \int_0^\infty h_{\alpha/2}(z) dz.$$

Relation (6) can be written in a somewhat different form in terms of the density $w_{\alpha/2}$ with the account of (3):

$$\ell_\alpha(x) = \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \int_0^\infty \frac{1}{\sqrt{2\pi z}} \exp\left\{-\frac{x^2}{2z}\right\} \frac{g_{\alpha/2,1}(z^{-1})}{z^{3/2}} dz = \int_0^\infty \frac{1}{\sqrt{z}} \varphi\left(\frac{x}{\sqrt{z}}\right) w_{\alpha/2}(z) dz. \quad (10)$$

Thus, representations (4) and (5) follow from (9) and (10), correspondingly. The lemma is proved.

If Z_α is a random variable having the exponential power distribution with parameter α , then relation (6) with the account of (3) mean that $Z_\alpha \stackrel{d}{=} X \cdot \sqrt{U_{\alpha/2}}$, where X and $U_{\alpha/2}$ are independent random variables such that X has the standard normal distribution.

Since the function $h_{\alpha/2}(x)$ is a probability density, then its definition implies the following interesting statement providing the possibility to calculate $EU_{\alpha/2,1}^{-1/2}$ explicitly although, in general, the density $g_{\alpha/2,1}(z)$ cannot be written out in an explicit form in terms of elementary or simple special functions. However, this statement has only a tangent relation to the main topic of this paper.

COROLLARY 1. *Let $Y_{\alpha,1}$ be a random variable having the one-sided stable distribution with characteristic exponent $\alpha \in (0, 1)$. Then $EY_{\alpha,1}^{-1/2} = \frac{1}{\alpha} \Gamma(\frac{1}{2\alpha}) / \sqrt{2\pi}$.*

EXAMPLE 1. Consider the case $\alpha = 1$. Then, as is known, $G_{1/2,1}(x)$ is the Lévy distribution (a particular case of the inverse Gaussian distribution, the distribution of the time until the standard Wiener process hits the unit level). The corresponding density is

$$g_{1/2,1}(z) = \frac{1}{z^{3/2} \sqrt{2\pi}} \exp\left\{-\frac{1}{2z}\right\}, \quad z > 0.$$

In this case

$$w_{1/2}(z) = \sqrt{\frac{\pi}{2}} \cdot \frac{g_{1/2,1}(z^{-1})}{z^{3/2}} = \frac{\sqrt{\pi} z^{3/2} e^{-z/2}}{\sqrt{2} \sqrt{2\pi} z^{3/2}} = \frac{1}{2} e^{-z/2},$$

that is, $w_{1/2}(z) = \frac{1}{2} e^{-z/2}$ is the density of the exponential distribution with parameter $\frac{1}{2}$. As this is so, according to (7) we have

$$\ell_1(x) = \frac{1}{2} e^{-|x|} = \int_0^\infty \frac{1}{\sqrt{z}} \varphi\left(\frac{x}{\sqrt{z}}\right) \frac{e^{-z/2}}{2} dz,$$

which is a well-known property of the Laplace distribution, see, e. g., (Korolev et al. 2011a), lemma 12.7.1.

A criterion of convergence of the distributions of random sums to exponential power distributions

Everywhere in what follows the symbol \implies denotes convergence in distribution.

Consider a sequence of independent identically distributed random variables X_1, X_2, \dots , defined on a probability space $(\Omega, \mathfrak{A}, P)$. Assume that

$$EX_1 = 0, \quad 0 < \sigma^2 = DX_1 < \infty. \quad (11)$$

For a natural $n \geq 1$ let $S_n = X_1 + \dots + X_n$. Let N_1, N_2, \dots be a sequence of nonnegative integer random variables defined on the same probability space so that for each $n \geq 1$ the random variable N_n is independent of the sequence X_1, X_2, \dots . For definiteness, hereinafter we will assume that $\sum_{j=1}^0 = 0$. In what follows convergence will mean as $n \rightarrow \infty$ unless otherwise specified.

A random sequence N_1, N_2, \dots is said to be infinitely increasing ($N_n \rightarrow \infty$) in probability, if $P(N_n \leq m) \rightarrow 0$ for any $m \in (0, \infty)$.

The following important statement was firstly proved in (Korolev 1994).

LEMMA 2. *Assume that the random variables X_1, X_2, \dots and N_1, N_2, \dots satisfy the conditions specified above and $N_n \rightarrow \infty$ in probability. A distribution function $F(x)$ such that $P(S_{N_n} < x\sigma\sqrt{n}) \implies F(x)$ exists if and only if there exists a distribution function $Q(x)$ satisfying the conditions $Q(0) = 0$,*

$$F(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{y}}\right) dQ(y), \quad x \in \mathbb{R}, \quad P(N_n < nx) \implies Q(x).$$

THEOREM 1. *Assume that the random variables X_1, X_2, \dots and N_1, N_2, \dots satisfy the conditions specified above and $N_n \rightarrow \infty$ in probability. Then $P(S_{N_n} < x\sigma\sqrt{n}) \implies L_\alpha(x)$, if and only if $P(N_n < nx) \implies W_{\alpha/2}(x)$.*

PROOF. This statement is a direct consequence of lemma 2 with $Q(x) = W_{\alpha/2}(x)$ and representation (5).

A criterion of convergence of the distributions of regular statistics constructed from samples with random sizes to exponential power distributions

For $n \geq 1$ let $T_n = T_n(X_1, \dots, X_n)$ be a statistic, that is, a measurable function of the random variables X_1, \dots, X_n . For each $n \geq 1$ define the random variable T_{N_n} by letting $T_{N_n}(\omega) = T_{N_n(\omega)}(X_1(\omega), \dots, X_{N_n(\omega)}(\omega))$ for every elementary outcome $\omega \in \Omega$.

We will say that the statistic T_n is asymptotically normal, if there exist $\delta > 0$ and $\theta \in \mathbb{R}$ such that

$$P(\delta\sqrt{n}(T_n - \theta) < x) \implies \Phi(x). \tag{12}$$

LEMMA 3. Assume that $N_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. Let the statistic T_n be asymptotically normal in the sense of (12). Then a distribution function $F(x)$ such that $P(\delta\sqrt{n}(T_{N_n} - \theta) < x) \implies F(x)$ exists if and only if there exists a distribution function $Q(x)$ satisfying the conditions $Q(0) = 0$, $F(x) = \int_0^\infty \Phi(x\sqrt{y})dQ(y)$, $x \in \mathbb{R}$, $P(N_n < nx) \implies Q(x)$.

PROOF. Actually, this lemma is a particular case of theorem 3 in (Korolev 1995), the proof of which is, in turn, based on general theorems on convergence of superpositions of independent random sequences (Korolev 1994, Korolev 1996). Also see (Gnedenko and Korolev 1996), theorem 3.3.2.

THEOREM 2. Assume that $N_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. Let the statistic T_n be asymptotically normal in the sense of (12). Then $P(\delta\sqrt{n}(T_{N_n} - \theta) < x) \implies L_\alpha(x)$ if and only if $P(N_n < nx) \implies H_{\alpha/2}(x)$.

PROOF. This statement is a direct consequence of lemma 3 with $Q(x) = H_{\alpha/2}(x)$ and representation (4).

Discussion

The convergence of the distributions of the normalized indices N_n to the distributions $W_{\alpha/2}$ and $H_{\alpha/2}$ is the main condition in theorems 1 and 2, respectively. Now we will give a rather general example of the situation where these conditions can hold. For this purpose we introduce a useful construction of nonnegative integer-valued random variables which, under an appropriate normalization, converge to a given nonnegative (not necessarily discrete) random variable, whatever the latter is.

In the book (Gnedenko and Korolev 1996) it was proposed to model the evolution of non-homogeneous chaotic stochastic processes, in particular, the dynamics of financial assets by compound doubly stochastic Poisson processes (compound Cox processes). This approach got further grounds and development in the books (Bening and Korolev 2002, Korolev and Sokolov

2008, Korolev et al. 2011a, Korolev 2011). In (Korolev and Skvortsova 2006, Korolev 2011) this approach was successfully applied to modeling the processes of plasma turbulence. Similar methods were considered in (Granovsky and Zeifman 2000, Zeifman 1991). According to this approach the flow of informative events, each of which generates the next observation, is described by the stochastic point process $M(\Lambda(t))$ where $M(t)$, $t \geq 0$, is a homogeneous Poisson process with unit intensity and $\Lambda(t)$, $t \geq 0$, is a random process independent of $M(t)$ possessing the properties: $\Lambda(0) = 0$, $P(\Lambda(t) < \infty) = 1$ for any $t > 0$, the trajectories $\Lambda(t)$ are non-decreasing and right-continuous. The process $M(\Lambda(t))$, $t \geq 0$, is called the doubly stochastic Poisson process (Cox process).

Within this model for each t the distribution of the random variable $M(\Lambda(t))$ is mixed Poisson. Consider the case where in this model the time t remains fixed (say, $t = 1$) and $\Lambda(t) = nU_{\alpha/2}$, where n is an auxiliary natural-valued parameter, $U_{\alpha/2}$ is a random variable with the distribution function $W_{\alpha/2}(x)$ independent of the standard Poisson process $M(t)$, $t \geq 0$. Here the asymptotic $n \rightarrow \infty$ can be interpreted as that the (stochastic) intensity of the flow of informative events is assumed very large. For each natural n let

$$N_n = M(nU_{\alpha/2}) \tag{13}.$$

It is obvious that the random variable N_n so defined has the mixed Poisson distribution

$$P(N_n = k) = P(M(nU_{\alpha/2}) = k) = \int_0^\infty e^{-nz} \frac{(nz)^k}{k!} w_{\alpha/2}(z) dz \quad k = 0, 1, \dots$$

This random variable N_n can be also interpreted as the number of events registered up to time n in the Poisson process with the stochastic intensity having the density $w_{\alpha/2}(z)$. Assume that the random variable $U_{\alpha/2}$ and the Poisson process $M(t)$ are independent of the sequence X_1, X_2, \dots . Then, obviously, for each n the random variable N_n is also independent of this sequence.

Denote $A_n(z) = P(N_n < nz)$, $z \geq 0$ ($A_n(z) = 0$ for $z < 0$). It is easy to see that $A_n(z) \implies W_{\alpha/2}(z)$. Indeed, as is known, if $\Pi(x; \ell)$ is the Poisson distribution function with the parameter $\ell > 0$ and $E(x; c)$ is the distribution function with a single unit jump at the point $c \in \mathbb{R}$, then $\Pi(\ell x; \ell) \implies E(x; 1)$ as $\ell \rightarrow \infty$. Since for $x \in \mathbb{R}$ $A_n(x) = \int_0^\infty \Pi(nx; nz) dW_{\alpha/2}(z)$, then by the Lebesgue dominated convergence theorem, as $n \rightarrow \infty$, we have

$$A_n(x) \implies \int_0^\infty E(x/z; 1) dW_{\alpha/2}(z) = \int_0^x dW_{\alpha/2}(z) = W_{\alpha/2}(x),$$

that is, the random variables N_n defined above satisfy the condition of lemma 2 with $Q(x) = W_{\alpha/2}(x)$. Moreover, $N_n \rightarrow \infty$ in probability since $P(U_{\alpha/2} = 0) = 0$.

Similarly, let $V_{\alpha/2}$ be a random variable with the distribution function $H_{\alpha/2}(x)$ independent of the standard

Poisson process $M(t)$, $t \geq 0$. For each natural n let $N_n = M(nV_{\alpha/2})$. The distribution of the random variable N_n is mixed Poisson,

$$P(N_n = k) = \frac{1}{k!} \int_0^\infty e^{-nz} (nz)^k h_{\alpha/2}(z) dz, \quad k = 0, 1, 2, \dots$$

This random variable N_n can be also interpreted as the number of events registered up to time n in the Poisson process with the stochastic intensity having the density $h_{\alpha/2}(z)$. Assume that the random variable $V_{\alpha/2}$ and the Poisson process $M(t)$ are independent of the sequence X_1, X_2, \dots . Then, obviously, for each n the random variable N_n is also independent of this sequence.

As above, it is easy to make sure that $P(N_n < nz) \implies H_{\alpha/2}(z)$, that is, these random variables N_n satisfy the condition of lemma 3 with $Q(x) = H_{\alpha/2}(x)$. Moreover, $N_n \rightarrow \infty$ in probability since $P(V_{\alpha/2} = 0) = 0$.

REMARK 1. Using this simple construction of random indices N_n through the random change of time in the Poisson process one can easily obtain examples of random variables N_n participating in lemmas 2 and 3 whatever a distribution function $Q(x)$ with $Q(0) = 0$ is. Indeed, if U is a random variable with the distribution function $Q(x)$ and $M(t)$, $t \geq 0$, is the standard Poisson process independent of U , then the random variables $N_n = M(nU)$ satisfy the condition $P(N_n < nx) \implies Q(x)$.

Estimates on the rate of convergence of the distributions of random sums to exponential power laws

Here we will consider the rate of convergence in theorem 1. In addition to the conditions on the random variables X_1, X_2, \dots imposed in Sect. 2, assume that

$$\beta^3 = E|X_1|^3 < \infty. \quad (14)$$

Let the random variable N_n be defined by (13). Denote $D_{n,\alpha} = \sup_x |P(S_{N_n} < x\sigma\sqrt{n}) - L_\alpha(x)|$.

THEOREM 3. *Let conditions (11) and (14) hold and let the random variable N_n be defined by (13). For any $n \geq 1$ we have*

$$D_{n,\alpha} \leq 0.3812 \cdot \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \frac{\beta^3}{\sigma^3 \sqrt{n}}.$$

PROOF. The distribution of the random variable N_n is mixed Poisson. Hence, by the Fubini theorem

$$\begin{aligned} P(S_{N_n} < x\sigma\sqrt{n}) &= P(S_{M(nV_{\alpha/2,1})} < x\sigma\sqrt{n}) = \\ &= \int_0^\infty P(S_{M(nz)} < x\sigma\sqrt{n}) w_{\alpha/2}(z) dz. \end{aligned} \quad (15)$$

Further, according to (5), the exponential power distribution with parameter α is a scale mixture of normal

laws in which the mixing distribution is $W_{\alpha/2}$. From (15) and (5) it follows that

$$\begin{aligned} D_{n,\alpha} &\leq \int_0^\infty \sup_x \left| P\left(\frac{S_{M(nz)}}{\sigma\sqrt{n}} < x\right) - \Phi\left(\frac{x}{\sqrt{z}}\right) \right| dW_{\alpha/2}(z) = \\ &= \int_0^\infty \sup_x \left| P\left(\frac{S_{M(nz)}}{\sigma\sqrt{nz}} < x\right) - \Phi(x) \right| dW_{\alpha/2}(z). \end{aligned} \quad (16)$$

For the estimation of the integrand in (16) we will use the following analog of the Berry–Esseen inequality for Poisson random sums in terms of non-central Lyapunov fractions.

LEMMA 4. *Let random variables X_1, X_2, \dots be identically distributed with $EX_1 = 0$ and $E|X_1|^3 < \infty$. Let M_λ be a Poisson random variable with parameter $\lambda > 0$ such that the random variables $M_\lambda, X_1, X_2, \dots$ are jointly independent. Denote $Z_\lambda = X_1 + \dots + X_{M_\lambda}$. Then*

$$\sup_x |P(Z_\lambda < x\sqrt{DZ_\lambda} < x) - \Phi(x)| \leq \frac{0.3041}{\sqrt{\lambda}} \cdot \frac{E|X_1|^3}{(EX_1^2)^{3/2}}.$$

The PROOF of this statement was given in (Korolev and Shevtsova 2012), also see (Korolev et al. 2011a), theorem 2.4.3.

We will also use the following statement which makes it possible to calculate $EU_{\alpha/2}^{-1/2}$ although, in general, the density $w_{\alpha/2}(z)$ cannot be written out in an explicit form in terms of elementary or simple special functions.

LEMMA 5. *For any $\alpha \in (0, 2)$ we have $EU_{\alpha/2}^{-1/2} = \frac{\alpha}{2} \sqrt{\pi} / \Gamma(\frac{1}{\alpha})$.*

PROOF. Obviously, $EU_{\alpha/2}^{-1/2} = EV_{\alpha/2}^{1/2} = \int_0^\infty \sqrt{z} h_{\alpha/2}(z) dz$. Further, from the definition of the density $h_{\alpha/2}$ it follows that

$$\begin{aligned} \int_0^\infty \sqrt{z} h_{\alpha/2}(z) dz &= \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \int_0^\infty \frac{\sqrt{z}}{\sqrt{z}} g_{\alpha/2,1}(z) dz = \\ &= \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \int_0^\infty g_{\alpha/2,1}(z) dz = \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}}, \end{aligned}$$

since $g_{\alpha/2,1}(z)$ is a probability density. The lemma is proved.

Continuing (16) with the account of lemmas 4 and 5 we obtain

$$D_{n,\alpha} \leq 0.3041 \cdot \frac{\beta^3}{\sigma^3 \sqrt{n}} \cdot EU_{\alpha/2}^{-1/2} = 0.3041 \cdot \frac{\alpha}{\Gamma(\frac{1}{\alpha})} \sqrt{\frac{\pi}{2}} \cdot \frac{\beta^3}{\sigma^3 \sqrt{n}}.$$

The theorem is proved.

Asymmetric generalization of exponential power distributions by variance-mean mixing

All the distributions of type (1) are symmetric. There were some attempts to propose asymmetric (skew) generalization of distributions (1), see, e. g. (Fernandez et al. 1995), (Theodossiou 2000), (Komunjer 2007) where

the so-called *skew exponential power distributions* were considered. In (Ayebo and Kozubowski 2004) basic properties of the skew exponential power distributions were considered. In (Zhu and Zinde-Walsh 2009) the so-called *asymmetric exponential power distributions* were proposed. However, all these generalizations are rather formal and do not assume the property of a "generalized" distribution to be a limit law in some simple asymptotic setting.

Instead, here we consider a more natural asymmetric extension of the class of distributions (1). For this purpose we will use an approach similar to the one used by O. Barndorff-Nielsen in 1977 to introduce the class of generalized hyperbolic distributions as special variance-mean mixtures of normal laws (Barndorff-Nielsen 1977). The base of the corresponding reasoning is representation (5).

Let $\alpha \in (0, 2]$, $\mu \in \mathbb{R}$. The probability distribution whose distribution function has the form

$$L_{\alpha, \mu}(x) = \int_0^\infty \Phi\left(\frac{x - \mu z}{\sqrt{z}}\right) dW_{\alpha/2}(z), \quad x \in \mathbb{R}, \quad (17)$$

will be called a *skew exponential power distribution* or *skew generalized Laplace distribution* with shape parameter α and asymmetry parameter μ . Formally, in mixture (17) the mixing is carried out with respect to both parameters of the normal law. However, by virtue of the fact that in (17) these parameters are tightly linked and the expectations (means) of the mixed normal laws turn out to be proportional to their variances, actually, (17) is a one-parameter mixture. That is why O. Barndorff-Nielsen and his colleagues called such mixtures *variance-mean mixtures* (Barndorff-Nielsen et al 1982).

If X is a random variable with the standard normal law independent of the random variable $U_{\alpha/2}$ introduced above, then it is easy to see that the distribution function $L_{\alpha, \mu}(x)$ (see (17)) corresponds to the random variable $Z_{\alpha, \mu} = X\sqrt{U_{\alpha/2}} + \mu U_{\alpha/2}$.

The moments of $Z_{\alpha, \mu}$ were found in (Grigoryeva and Korolev 2013).

A criterion of convergence of the distributions of random sums to skew exponential power distributions

Let $\{X_{n,j}\}_{j \geq 1}$, $n = 1, 2, \dots$ be a double array of row-wise identically distributed random variables. Let $\{N_n\}_{n \geq 1}$ be a sequence of integer-valued nonnegative random variables such that for each $n \geq 1$ the random variables $N_n, X_{n,1}, X_{n,2}, \dots$ are independent. Let $S_{n,k} = X_{n,1} + \dots + X_{n,k}$. As above, to avoid misunderstanding we assume $\sum_{j=1}^0 = 0$.

As it was demonstrated in (Korolev 2013, Zaks and Korolev 2013), variance-mean mixtures of normal laws (3) turn out to be identifiable, since for each fixed $\mu \in$

\mathbb{R} the one-parameter family of distributions $\{\Phi((x - \mu z)/\sqrt{z}) : z \geq 0\}$ is additively closed. In (Korolev 2013) the following general statement was proved (also see (Zaks and Korolev 2013)).

LEMMA 6. Assume that there exist a sequence $\{k_n\}_{n \geq 1}$ of natural numbers and a number $\mu \in \mathbb{R}$ such that

$$P(S_{n,k_n} < x) \implies \Phi(x - \mu). \quad (18)$$

Assume that $N_n \rightarrow \infty$ in probability. Then the distributions of random sums weakly converge to some distribution function $F(x) : P(S_{n,N_n} < x) \implies F(x)$, if and only if there exists a distribution function $Q(x)$ such that $Q(0) = 0$,

$$F(x) = \int_0^\infty \Phi\left(\frac{x - \mu z}{\sqrt{z}}\right) dQ(z), \quad (19)$$

and

$$P(N_n < x k_n) \implies Q(x). \quad (20)$$

The following theorem is actually a particular case of lemma 6.

THEOREM 4. Assume that there exist a sequence $\{k_n\}_{n \geq 1}$ of natural numbers and a number $\mu \in \mathbb{R}$ such that convergence (6) takes place. Assume that $N_n \rightarrow \infty$ in probability. Then convergence

$$P(S_{n,N_n} < x) \implies L_{\alpha, \mu}(x) \quad (21)$$

takes place if and only if

$$P(N_n < x k_n) \implies W_{\alpha/2}(x). \quad (22)$$

Some estimates of the rate of convergence in theorem 4 were presented in (Grigoryeva and Korolev 2013).

Research supported by the Russian Foundation for Basic Research (projects 12-07-00115a, 12-07-00109a, 14-07-00041a).

REFERENCES

- Ayebo A., T.J. Kozubowski. 2004. "An asymmetric generalization of Gaussian and Laplace laws." *Journal of Probability and Statistical Science*. 1(2), 187–210.
- Barndorff-Nielsen O. E. 1977. "Exponentially decreasing distributions for the logarithm of particle size." *Proc. Roy. Soc. London, Ser. A*. A(353), 401–419.
- Barndorff-Nielsen O. E., J. Kent and M. Sørensen. 1982. "Normal variance-mean mixtures and z-distributions." *International Statistical Review*. 50(2), 145–159.
- Bening V., V. Korolev V. 2002. *Generalized Poisson Models and their Applications in Insurance and Finance*. VSP, Utrecht.
- Box G., G. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. Wiley, New York.

- Choy S. T. B., A. F. F. Smith. 1997. "Hierarchical models with scale mixtures of normal distributions." *TEST*. 6, 205–221.
- Evans M., N. Hastings and B. Peacock. 2000. *Statistical Distributions*, 3rd. edn. Wiley, New York.
- Fernandez C., J. Osiewalski and M.F.J. Steel. 1995. "Modeling and inference with v -distributions." *J. Amer. Statist. Assoc.* 90(432), 1331–1340.
- Gnedenko B. V., A.N. Kolmogorov. 1954. *Limit Distributions Sums of Independent Random Variables*. Addison-Wesley, Cambridge, MA.
- Gnedenko B. V., V. Yu. Korolev. 1996. *Random Summation: Limit Theorems and Applications*. CRC Press, Boca Raton.
- Granovsky B. L., A.I. Zeifman. 2000. "The N-limit of spectral gap of a class of birth-death Markov chains." *Applied Stochastic Models in Business and Industry*. 16, 235–248.
- Grigoryeva M. E., V.Yu. Korolev. 2013. "On convergence of the distributions of random sums to skew exponential power distributions." *Informatics and Its Applications*. 7(4), 66–74.
- Komunjer I. 2007. "Asymmetric power distribution: Theory and applications to risk measurement." *Journal of Applied Econometrics*. 22, 891–921.
- Korolev V. Yu. 1994. "Convergence of random sequences with independent random indices. I." *Theory Probab. Appl.* 39, 313–333.
- Korolev V. Yu. 1995. "Convergence of random sequences with independent random indices. II." *Theory Probab. Appl.* 40, 907–910.
- Korolev V. Yu. 1996. "A general theorem on the limit behavior of superpositions of independent random processes with applications to Cox processes." *Journal of Mathematical Sciences*. 81, 2951–2956.
- Korolev V., N. Skvortsova (Eds.). 2006. *Stochastic Models of Structural Plasma Turbulence*. VSP, Utrecht.
- Korolev V. Yu., I.A. Sokolov. 2008. *Mathematical Models of Non-homogeneous Flows of Extremal Events*. Torus, Moscow (in Russian).
- Korolev V. Yu. 2011. *Probabilistic and Statistical Methods of Decomposition of Volatility of Chaotic Processes*. Moscow University Publishing House, Moscow (in Russian).
- Korolev V. Yu., V.E. Bening and S.Ya. Shorgin. 2011a. *Mathematical Foundations of Risk Theory*, 2nd ed. FIZMATLIT, Moscow (in Russian).
- Korolev V. Yu., I.G. Shevtsova and S.Ya. Shorgin. 2011b. "On Berry–Esseen-type inequalities for Poisson random sums." *Informatica and Its Applications*. 5(3), 64–66.
- Korolev V., I. Shevtsova I. 2012. "An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums". *Scandinavian Actuarial Journal*. 12, no. 2, 81–105.
- Korolev V. Yu. 2013. "Generalized hyperbolic distributions as limit laws for random sums." *Theory Probab. Appl.* 57(1), 117–132.
- Leemis L. M., J.T. McQueston. 2008. "Univariate distribution relationships." *The American Statistician*. 62, no. 1, 45–53.
- Nadaraja S. 2005. "A generalized normal distribution." *Journal of Applied Statistics*. 32, 685–694.
- Morgan J.P. 1996. *RiskMetrics Technical Document*. Risk-Metric Group, New York.
- Subbotin M. T. 1923. "On the law of frequency of error." *Matematicheskii Sbornik*. 31, 296–301.
- Theodossiou P. 2000. *Skewed Generalized Error Distribution of Financial Assets and Option Pricing*. SSRN working paper.
- Varanasi M. K., B. Aazhang. 1989. "Parametric generalized Gaussian density estimation." *Journal of the Acoustic Society of America*. 86, 1404–1415.
- West M. 1987. "On scale mixtures of normal distributions." *Biometrika*. 74, 646–648.
- Zaks L. M., V.Yu. Korolev. 2013. "Generalized variance-gamma distributions as limit laws for random sums." *Informatics and Its Applications*. 7(1), 105–115.
- Zeifman A. I. 1991. "Some estimates of the rate of convergence for birth and death processes." *Journal of Applied Probability*. 28, 268–277.
- Zhu D., V. Zinde-Walsh. 2009. "Properties and estimation of asymmetric exponential power distribution." *Journal of Econometrics*. 148(1), 86–99.
- Zolotarev V. M. 1986. *One-dimensional Stable Distributions*. American Mathematical Society, Providence, RI.

AUTHOR BIOGRAPHIES

VICTOR KOROLEV is Doctor of Science in physics and mathematics, professor, Head of Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

MARIA GRIGORYEVA is biostatistic II at Parexel International, Moscow. Her email is maria-grigoryeva@yandex.ru.

ALEXANDER ZEIFMAN Doctor of Science in physics and mathematics; professor, Head of Department of Applied Mathematics, Vologda State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; principal scientist, Institute of Territories Socio-Economic Development, Russian Academy of Sciences. His email is a.zeifman@mail.ru and his personal webpage at <http://uni-vologda.ac.ru/zai/eng.html>.

TRANSFER THEOREM CONCERNING ASYMPTOTIC EXPANSIONS FOR THE DISTRIBUTION FUNCTIONS OF STATISTICS BASED ON SAMPLES WITH RANDOM SIZES

Vladimir Bening

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University
Institute for Informatics Problems, Russian Academy of Sciences

V. A. Savushkin

Dubna State University

E. I. Shunkov

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University

Alexander Zeifman

Vologda State University, Russia

Institute of Informatics Problems, Russian Academy of Sciences

Institute of Territories Socio-Economic Development, Russian Academy of Sciences

Victor Korolev

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University

Institute for Informatics Problems, Russian Academy of Sciences

KEYWORDS

Random size; asymptotic expansion; transfer theorem; Student distribution; Laplace distribution

ABSTRACT

In the paper, we discuss the transformation of the asymptotic expansion for the distribution of a statistic admitting Edgeworth expansion if the sample size is replaced by a random variable. We demonstrate that all those statistics that are regarded as asymptotically normal in the classical sense, become asymptotically Laplace or Student if the sample size is random. Thus, the Laplace and Student distributions may be used as an asymptotic approximation in descriptive statistics being a convenient heavy-tailed alternative to stable laws.

INTRODUCTION

In 1774 P. S. Laplace in his paper "Sur la probabilité des causes par les événements" (see (Kotz et al. 2001) and references in the book) introduced a native probabilistic law for the error of measurement in the following formulation: "the logarithm of the frequency of an error (without regard to sign) is a linear function of the error". Later in 1911 the famous economist and probabilist J. M. Keynes obtained the first law error again from the assumption that the most probable value of the measured quantity is equal to the median of measurements (see (Kotz et al. 2001) and references in the book).

Later in 1923 E. B. Wilson suggested that the frequency we actually meet in everyday work in economics, biometrics, or vital statistics often fails to conform closely to the normal distribution, and that Laplace's first law should be considered as a candidate for fitting data in economics and health sciences (see (Kotz et al. 2001) and references in the book). Fifty years later in scientific papers (see (Kotz et al. 2001) and references in the book) one could often find appeals for using the first Laplace's law as the main hypothesis instead of the normal distribution for the economical, biometrical and demographic data.

Nowadays the first Laplace's law is called the Laplace distribution. The distribution is defined by its characteristic function (see (Bening and Korolev 2008) and the references therein)

$$f(s) = \frac{2}{2 + \nu^2 s^2}, \quad s \in \mathbb{R}^1, \quad (1.1)$$

or by its density

$$l(x) = \frac{1}{\nu\sqrt{2}} \exp\left\{-\frac{\sqrt{2}|x|}{\nu}\right\}, \quad \nu > 0, \quad x \in \mathbb{R}^1. \quad (1.2)$$

Another name – double exponential distribution – shows an opportunity to obtain it as the difference between two independent identically distributed exponential random variables which are often used for modeling of lifetime of an observable object.

We now present the reasoning from (Bening and Korolev 2008) which validates the use of Laplace distribution in problems of probability theory and mathe-

mathematical statistics as the limiting distribution for samples of random size. Consider random variables $N_1, N_2, \dots, X_1, X_2, \dots$ defined on a common measurable space (Ω, \mathcal{A}) . Let P be a probability measure over (Ω, \mathcal{A}) . Suppose that the random variables N_n take on positive integers for any $n \geq 1$ and do not depend on X_1, X_2, \dots . Define the random variable T_{N_n} for some statistic $T_n = T_n(X_1, \dots, X_n)$ and any $n \geq 1$ by

$$T_{N_n}(\omega) = T_{N_n(\omega)}(X_1(\omega), \dots, X_{N_n(\omega)}(\omega)),$$

for every outcome $\omega \in \Omega$. The statistic T_n is called asymptotically normal if there exist real numbers $\sigma > 0$ and $\mu \in \mathbb{R}^1$ such that, as $n \rightarrow \infty$,

$$P(\sigma\sqrt{n}(T_n - \mu) < x) \implies \Phi(x), \quad (1.3)$$

where $\Phi(x)$ is the standard normal distribution function.

The asymptotically normal statistics are abundant. Paper (Bening and Korolev 2008) contains some examples of these statistics: the sample mean (assuming nonzero variances), the central order statistics or the maximum likelihood estimators (under weak regularity conditions) and many others. The following lemma, proved in (Bening and Korolev 2008), gives the necessary and sufficient conditions under which the distributions of asymptotically normal statistics based on samples of random size converge to a predetermined distribution $F(x)$.

Lemma 1.1. (Korolev 1995) *Let $\{d_n\}_{n \geq 1}$ be an increasing and unbounded sequence of positive numbers. Suppose that $N_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. Let T_n be an asymptotically normal statistic as in (1.3). Then a necessary and sufficient condition for a distribution function $F(x)$ to satisfy*

$$P(\sigma\sqrt{d_n}(T_{N_n} - \mu) < x) \implies F(x) \quad (n \rightarrow \infty)$$

is that there exists a distribution function $H(x)$ satisfying

$$H(x) = 0, \quad x < 0;$$

$$F(x) = \int_0^\infty \Phi(x\sqrt{y})dH(y), \quad x \in \mathbb{R}^1;$$

$$P(N_n < d_n x) \implies H(x) \quad (n \rightarrow \infty).$$

It is well known (see e.g. (Bening and Korolev 2008)) that the Laplace distribution can be expressed in terms of a scale mixture of normal distributions (with zero mean) with an inverse exponential mixing distribution, i.e., for any $x \in \mathbb{R}^1$,

$$L(x) = \int_0^\infty \Phi(x\sqrt{y})dQ(y),$$

where $Q(x)$ is the distribution function of the inverse exponential distribution

$$Q(x) = e^{-\delta/x}, \quad \delta > 0, \quad x > 0,$$

and $L(x)$ is the distribution function of the Laplace distribution corresponding to the density (1.2) with $\nu^2 = 1/\delta$.

Recall that the inverse exponential distribution is the distribution of the random variable

$$V = \frac{1}{U},$$

where the random variable U has the exponential distribution, and the inverse exponential distribution is a special case of the Fréchet distribution which is well known in asymptotic theory of order statistics as the type II extreme value distribution.

Lemma 1.1 can be applied to derive the following theorem which gives the necessary and sufficient conditions for the Laplace distribution to be the limiting distribution of the asymptotically normal statistics based on samples of random size.

Theorem 1.2. (Bening and Korolev 2008) *Let $\sigma > 0$ and $\{d_n\}_{n \geq 1}$ be an increasing and unbounded sequence of positive numbers. Suppose that $N_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. Let T_n be an asymptotically normal statistic as in (1.3). Then*

$$P(\sigma\sqrt{d_n}(T_{N_n} - \mu) < x) \implies L(x) \quad (n \rightarrow \infty)$$

if and only if

$$P(N_n < d_n x) \implies Q(x) \quad (n \rightarrow \infty).$$

Consider an example from (Bening and Korolev 2008) in which the random size of sample has the limiting inverse exponential distribution $Q(x)$. Let Y_1, Y_2, \dots be the independent and identically distributed random variables with some continuous distribution function. Let m be a positive integer and

$$N(m) = \min\{n \geq 1 : \max_{1 \leq j \leq m} Y_j < \max_{m+1 \leq k \leq m+n} Y_k\}.$$

The random variable $N(m)$ denotes the number of additional observations needed to exceed the current maximum obtained with m observations. The distribution of the random variable $N(m)$ was obtained by S.S. Wilks (Wilks 1959). So, the distribution of $N(m)$ is the discrete Pareto distribution

$$P(N(m) \geq k) = \frac{m}{m+k-1}, \quad k \geq 1. \quad (1.4)$$

Now, let $N^{(1)}(m), N^{(2)}(m), \dots$ be the independent random variables with the same distribution (1.4). Then the following statement was proved in (Bening and Korolev 2008): for any $x > 0$,

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{n} \max_{1 \leq j \leq n} N^{(j)}(m) < x\right) = e^{-m/x}.$$

Therefore, the limit is the distribution function of the inverse exponential distribution with $\delta = m$. And if

$$N_n = \max_{1 \leq j \leq n} N^{(j)}(m), \quad (1.5)$$

then Theorem 1.2 (with $d_n = n$) gives the Laplace distribution as the limiting distribution of regular statistics.

Theorem 1.3. (Bening and Korolev 2008)
Let m be any positive integer. Suppose that $N^{(1)}(m), N^{(2)}(m), \dots$ are independent random variables having the same distribution (1.4), and a random variable N_n is defined by (1.5). Let T_n be an asymptotically normal statistic as in (1.3). Then

$$P(\sigma\sqrt{n}(T_{N_n} - \mu) < x) \implies L(x) \quad (n \rightarrow \infty),$$

where $L(x)$ is the distribution function of the Laplace distribution with density (1.2) with $\nu^2 = 1/m$.

Further, the Laplace distribution plays the same role in the theory of geometric random sums as the normal distribution plays in the classical probability theory (see e.g. (Bening and Korolev 2008) and the references therein). In turn, the geometric random sums play an important role in the investigation of speculative processes. The reason of increasing usage of the Laplace distribution is also its representation as a scale mixture of some well known distributions. For example, the Laplace distribution can be represented as a scale mixture of symmetrized Rayleigh-Rice distribution with the mixing χ^2 -distribution with 1 degree of freedom (see Corollary 3.2 in (Bening and Korolev 2008)).

The Laplace distribution as a probabilistic model for applications is also attractive because of its extremal entropy property. This property often motivates a choice of Laplace distribution as a model for the error of measurements when the accuracy randomly varies from one measurement to the next (see (Bening and Korolev 2008)).

In applied economics and science, the popularity of Laplace distribution as a mathematical (probabilistic) model is explained by the fact that the Laplace distribution has heavier tails than the normal distribution does. So, in communication theory, the Laplace distribution is considered as a probabilistic model for some types of random noise in problems of detection of a known constant signal (see (Astrabadi 1985, Dadi and Marks 1987, Marks et al. 1978, Miller and Thomas 1972). In (Duttweiler and Messerschmitt 1976) the Laplace distribution is referred to as a model for speech signal in problems of encoding and decoding of analog signals. In (Epstein 1948) an application of the Laplace distribution is discussed in relation to the fracturing of materials under applied forces. In (Jones and McLachlan 1990, Kanji 1985) authors give examples of application of Laplace distribution in aerodynamics, when the gradient of airspeed change against its duration is modeled by mixtures of the Laplace distribution with the normal distribution. Modeling of the error distributions in navigation with Laplace distribution is investigated in (Hsu 1979).

This increased interest in Laplace distribution from applied sciences motivates the Laplace distribution to

be investigated in mathematical statistics and theory of probability. The non-regularity of the Laplace distribution makes known difficulties of its use in problems of testing statistical hypotheses. But the asymptotic methods of testing statistical hypotheses developed in last decades now allow to use the Laplace distribution in mathematical statistics (see (Kotz et al. 2001) and references in the work).

ASYMPTOTIC EXPANSIONS

Consider random variables (r.v.'s) N_1, N_2, \dots and X_1, X_2, \dots , defined on the same probability space (Ω, \mathcal{A}, P) . By X_1, X_2, \dots, X_n we will mean statistical observations whereas the r.v. N_n will be regarded as the random sample size depending on the parameter $n \in \mathbb{N}$. Assume that for each $n \geq 1$ the r.v. N_n takes only natural values (i.e., $N_n \in \mathbb{N}$) and is independent of the sequence X_1, X_2, \dots . Everywhere in what follows the r.v.'s X_1, X_2, \dots are assumed independent and identically distributed.

For every $n \geq 1$ by $T_n = T_n(X_1, \dots, X_n)$ denote a statistic, i.e., a real-valued measurable function of X_1, \dots, X_n . For each $n \geq 1$ we define a r.v. T_{N_n} by setting $T_{N_n}(\omega) \equiv T_{N_n(\omega)}(X_1(\omega), \dots, X_{N_n(\omega)}(\omega))$, $\omega \in \Omega$.

The following condition determines the asymptotic expansion (a.e.) for the distribution function (d.f.) of T_n with a non-random sample size.

Condition 1. *There exist $l \in \mathbb{N}$, $\mu \in \mathbb{R}$, $\sigma > 0$, $\alpha > l/2$, $\gamma \geq 0$, $C_1 > 0$, a differentiable d.f. $F(x)$ and differentiable bounded functions $f_j(x)$, $j = 1, \dots, l$ such that*

$$\sup_x |P(\sigma n^\gamma (T_n - \mu) < x) - F(x) - \sum_{j=1}^l n^{-j/2} f_j(x)| \leq \frac{C_1}{n^\alpha}, \quad n \in \mathbb{N}.$$

The following condition determines the a.e. for the d.f. of the normalized random index N_n .

Condition 2. *There exist $m \in \mathbb{N}$, $\beta > m/2$, $C_2 > 0$, a function $0 < g(n) \uparrow \infty, n \rightarrow \infty$, a d.f. $H(x)$, $H(0+) = 0$ and functions $h_i(x)$, $i = 1, \dots, m$ with bounded variation such that*

$$\sup_{x \geq 0} |P\left(\frac{N_n}{g(n)} < x\right) - H(x) - \sum_{i=1}^m n^{-i/2} h_i(x)| \leq \frac{C_2}{n^\beta}, \quad n \in \mathbb{N}.$$

Define the function $G_n(x)$ as

$$\begin{aligned} G_n(x) = & \int_{1/g(n)}^{\infty} F(xy^\gamma) dH(y) + \sum_{i=1}^m n^{-i/2} \int_{1/g(n)}^{\infty} F(xy^\gamma) dh_i(y) + \\ & + \sum_{j=1}^l g^{-j/2}(n) \int_{1/g(n)}^{\infty} y^{-j/2} f_j(xy^\gamma) dH(y) + \\ & + \sum_{j=1}^l \sum_{i=1}^m n^{-i/2} g^{-j/2}(n) \int_{1/g(n)}^{\infty} y^{-j/2} f_j(xy^\gamma) dh_i(y). \end{aligned} \quad (2.1)$$

Theorem 2.1. Let the statistic $T_n = T_n(X_1, \dots, X_n)$ satisfy Condition 1 and the r.v. N_n satisfy Condition 2. Then there exists a constant $C_3 > 0$ such that

$$\sup_x |\mathbb{P}(\sigma g^\gamma(n)(T_{N_n} - \mu) < x) - G_n(x)| \leq C_1 \mathbb{E}N_n^{-\alpha} + \frac{C_3 + C_2 M_n}{n^\beta},$$

where

$$M_n = \sup_x \int_{1/g(n)}^\infty \left| \frac{\partial}{\partial y} (F(xy^\gamma) + \sum_{j=1}^l (yg(n))^{-j/2} f_j(xy^\gamma)) \right| dy \quad g_r(x) = \int_0^\infty \varphi(x\sqrt{y}) \frac{1-x^2y}{\sqrt{y}} dH_r(y), \quad x \geq 0. \quad (2.3)$$

and the function $G_n(x)$ is defined by (2.1).

Let $\Phi(x)$ and $\varphi(x)$ respectively denote the d.f. of the standard normal law and its density.

Lemma 2.1. Let $l = 1$, $0 < g(n) \uparrow \infty$, $F(x) = \Phi(x)$, $f_1(x) = \frac{1}{6}\mu_3\sigma^3(1-x^2)\varphi(x)$. Then the quantity M_n in Theorem 2.1 satisfies the inequality $M_n \leq 2 + \tilde{C}|\mu_3|\sigma^3$, where

$$\tilde{C} = \frac{1}{3} \sup_{u \geq 0} \{\varphi(u)(u^4 + 2u^2 + 1)\} = \frac{16}{3\sqrt{2\pi}e^3} \approx 0.47.$$

Consider some examples of application of Theorem 2.1.

Student distribution

Let X_1, X_2, \dots be i.i.d. r.v.'s with $\mathbb{E}X_1 = \mu$, $0 < \mathbb{D}X_1 = \sigma^{-2}$, $\mathbb{E}|X_1|^{3+2\delta} < \infty$, $\delta \in (0, \frac{1}{2})$ and $\mathbb{E}(X_1 - \mu)^3 = \mu_3$. For each n let

$$T_n = \frac{1}{n}(X_1 + \dots + X_n). \quad (2.2)$$

Assume that the r.v. X_1 satisfies the Cramér Condition (C)

$$\limsup_{|t| \rightarrow \infty} |\mathbb{E} \exp\{itX_1\}| < 1.$$

Let $G_\nu(x)$ be the Student d.f. with parameter $\nu > 0$ corresponding to the density

$$p_\nu(x) = \frac{\Gamma(\nu + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R},$$

where $\Gamma(\cdot)$ is the Euler's gamma-function and $\nu > 0$ is the shape parameter (if $\nu \in \mathbb{N}$, then ν is called the number of degrees of freedom). In practice, it can be arbitrarily small determining the typical heavy-tailed distribution. If $\nu = 2$, then the d.f. $G_2(x)$ is expressed explicitly as

$$G_2(x) = \frac{1}{2} \left(1 + \frac{x}{\sqrt{2+x^2}}\right), \quad x \in \mathbb{R}.$$

for $\nu = 1$ we have the Cauchy distribution.

For $r > 0$ let

$$H_r(x) = \frac{r^r}{\Gamma(r)} \int_0^x e^{-ry} y^{r-1} dy, \quad x \geq 0,$$

be the gamma-d.f. with parameter $r > 0$. Denote

$$g_r(x) = \int_0^\infty \varphi(x\sqrt{y}) \frac{1-x^2y}{\sqrt{y}} dH_r(y), \quad x \geq 0. \quad (2.3)$$

Theorem 2.2. Let the statistic T_n have the form (2.2), where X_1, X_2, \dots are i.i.d. r.v.'s with $\mathbb{E}X_1 = \mu$, $0 < \mathbb{D}X_1 = \sigma^{-2}$, $\mathbb{E}|X_1|^{3+2\delta} < \infty$, $\delta \in (0, \frac{1}{2})$ and $\mathbb{E}(X_1 - \mu)^3 = \mu_3$. Moreover, assume that the r.v. X_1 satisfies the Cramér Condition (C). Assume that for some $r > 0$ the r.v. N_n has the negative binomial distribution

$$\mathbb{P}(N_n = k) = \frac{(k+r-2) \cdots r}{(k-1)!} \frac{1}{n^r} \left(1 - \frac{1}{n}\right)^{k-1}, \quad k \in \mathbb{N}.$$

Let $G_{2r}(x)$ be the Student d.f. with parameter $\nu = 2r$ and $g_r(x)$ be defined by (2.3). Then for $r > 1/(1+2\delta)$, as $n \rightarrow \infty$, we have

$$\sup_x |\mathbb{P}(\sigma\sqrt{r(n-1)+1}(T_{N_n} - \mu) < x) - G_{2r}(x) - \frac{\mu_3\sigma^3 g_r(x)}{6\sqrt{r(n-1)+1}}| =$$

$$= \begin{cases} O\left(\left(\frac{\log n}{n}\right)^{1/2+\delta}\right), & r = 1, \\ O(n^{-\min(1, r(1/2+\delta))}), & r > 1, \\ O(n^{-r(1/2+\delta)}), & (1+2\delta)^{-1} < r < 1. \end{cases}$$

Laplace distribution

Consider the Laplace d.f. $\Lambda_\theta(x)$ corresponding to the density

$$\lambda_\theta(x) = \frac{1}{\theta\sqrt{2}} \exp\left\{-\frac{\sqrt{2}|x|}{\theta}\right\}, \quad \theta > 0, \quad x \in \mathbb{R}.$$

Let Y_1, Y_2, \dots be i.i.d. r.v.'s with a continuous d.f. Set

$$N(s) = \min\{i \geq 1 : \max_{1 \leq j \leq s} Y_j < \max_{s+1 \leq k \leq s+i} Y_k\}.$$

It is known that

$$\mathbb{P}(N(s) \geq k) = \frac{s}{s+k-1}, \quad k \geq 1 \quad (2.4)$$

(see, e.g., (Wilks 1959 or Nevzorov 2000)). Now let $N^{(1)}(s), N^{(2)}(s), \dots$ be i.i.d. r.v.'s distributed in accordance with (2.4). Define the r.v.

$$N_n(s) = \max_{1 \leq j \leq n} N^{(j)}(s),$$

then, as it was shown in (Bening and Korolev 2008),

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{N_n(s)}{n} < x\right) = e^{-s/x}, \quad x > 0,$$

and for an asymptotically normal statistic T_n we have

$$\mathbb{P}(\sigma\sqrt{n}(T_{N_n(s)} - \mu) < x) \longrightarrow \Lambda_{1/s}(x), \quad n \rightarrow \infty, \quad x \in \mathbb{R},$$

where $\Lambda_{1/s}(x)$ is the Laplace d.f. with parameter $\theta = 1/s$.

Denote

$$l_s(x) = \int_0^\infty \varphi(x\sqrt{y}) \frac{1 - x^2 y}{\sqrt{y}} de^{-s/y}, \quad x \in \mathbb{R}. \quad (2.5)$$

Theorem 2.3. *Let the statistic T_n have the form (2.2), where X_1, X_2, \dots are i.i.d. r.v.'s with $\mathbb{E}X_1 = \mu$, $0 < \mathbb{D}X_1 = \sigma^{-2}$, $\mathbb{E}|X_1|^{3+2\delta} < \infty$, $\delta \in (0, \frac{1}{2})$ and $\mathbb{E}(X_1 - \mu)^3 = \mu_3$. Moreover, assume that the r.v. X_1 satisfies the Cramér Condition (C). Assume that for some $s \in \mathbb{N}$ the r.v. $N_n(s)$ has the distribution*

$$\mathbb{P}(N_n(s) = k) = \binom{k}{s+k}^n - \binom{k-1}{s+k-1}^n, \quad k \in \mathbb{N}.$$

Then

$$\begin{aligned} \sup_x \left| \mathbb{P}(\sigma\sqrt{n}(T_{N_n(s)} - \mu) < x) - \Lambda_{1/s}(x) - \frac{\mu_3 \sigma^3 l_s(x)}{6\sqrt{n}} \right| = \\ = O\left(\frac{1}{n^{1/2+\delta}}\right), \quad n \rightarrow \infty, \end{aligned}$$

where $l_s(x)$ is defined in (2.5).

APPLICATION OF STUDENT DISTRIBUTION IN INSURANCE

In the 1960s, F. Bichsel suggested a risk rating system, called the Bonus-Malus system, which was better adjusted to the individual driver risk profiles. In the 1960s, car insurers requested approval for the increase of premium rates, claiming that the current level was insufficient to cover their risks. The supervision authority was prepared to give approval only if the rates took into account individual claims experience. It was no longer acceptable that "good" drivers, who had never made a claim, should continue to pay premiums which were at the same level as "bad" drivers who had made numerous claims.

Bichsel's Problem

Let N be the number of claims made by a particular driver in a year. The model used by Bichsel for the claim number is based on the following:

1. Conditionally, given $\Theta = \theta$, the N is Poisson distributed with Poisson parameter θ , i.e.

$$\mathbb{P}(N = k | \Theta = \theta) = e^{-\theta} \frac{\theta^k}{k!}, \quad k = 0, 1, \dots$$

2. Θ has a Gamma distribution with shape parameter r and a scale parameter β with the density

$$u(\theta) = \frac{\beta^r}{\Gamma(r)} \theta^{r-1} e^{-\beta\theta}, \quad \theta \geq 0.$$

The distribution function of Θ is called the structural function of the collective and describes the personal beliefs, a priori knowledge, and experience of the actuary.

The unconditional distribution of the number of claims is

$$\begin{aligned} \mathbb{P}(N = k) &= \int_0^\infty \mathbb{P}(N = k | \Theta = \theta) u(\theta) d\theta = \\ &= \int_0^\infty e^{-\theta} \frac{\theta^k}{k!} \frac{\beta^r}{\Gamma(r)} \theta^{r-1} e^{-\beta\theta} d\theta = \\ &= C_{r+k-1}^k p^r (1-p)^k, \quad k = 0, 1, \dots, \end{aligned}$$

where $p = \frac{\beta}{\beta+1}$, and $N \equiv N_{p,r}$ is the negative binomial random variable with parameters p and r .

Approximation of the Aggregate Claim Amount

Consider the statistic which is the average of claim amounts

$$T_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

where X_i is a claim size of each claim. Suppose that X_1, \dots, X_n are iid random variables, and $\mathbb{E}X_i = \mu$, $\mathbb{D}X_i = v^2$, $\sigma^2 = 1/v^2$. By CLT, we have

$$\mathbb{P}(\sigma\sqrt{n}(T_n - \mu) < x) \longrightarrow \Phi(x), \quad n \rightarrow \infty.$$

From our results we have an approximate formula for the aggregate claim amount for small β

$$\sum_{i=1}^{N_{p,r}} X_i \approx \frac{1}{\sigma} \sqrt{\frac{p}{r}} N_{p,r} S_{2r} + \mu,$$

where $p = \frac{\beta}{\beta+1} \approx 0$, and S_{2r} is the Student distributed random variable with parameter $2r$.

Research supported by the Russian Foundation for Basic Research (projects 12-07-00115a, 12-07-00109a, 14-07-00041a).

REFERENCES

- Asrabadi, B.R. 1985. "The exact confidence interval for the scale parameter and the MVUE of the Laplace distribution." *Communications in Statistics. Theory and Methods*. 14, 713–733.
- Bening, V.E. and V.Yu. Korolev. 2008. "Some statistical problems related to the Laplace distribution." *Informatics and its Applications*. 2(2), 19–34.
- Dadi, M.I. and R.J. Marks. 1987. "Detector relative efficiencies in the presence of Laplace noise." *IEEE Trans. Aerospace Electron. Systems*. 23(4), 568–582.
- Duttweiler, D.L. and D.G. Messerschmitt. 1976. "Nearly instantaneous companding for nonuniformly quantized PCM." *IEEE Trans. Comm.* 24(8), 864–873.
- Epstein, B. 1948. "Application of the theory of extreme values in fracture problems." *J. Amer. Statist. Assoc.* 43(243), 403–412.
- Hsu, D.A. 1979. "Long-tailed distributions for position errors in navigation." *Appl. Statist.* 28(1), 62–72.
- Jones, P.N. and G.J. McLachlan. 1990. "Laplace-normal mixtures fitted to wind shear data." *J. Appl. Statistics*. 17(2), 271–276.
- Kanji, G.K. 1985. "A mixture model for wind shear data." *J. Appl. Statistics*. 12(1), 49–58.
- Korolev, V.Yu. 1995. "Convergence of random sequences with independent random indices. II." *Theory of Probability and Its Applications*. 40(4), 770–772.
- Kotz, S., T.J. Kozubowski and K. Podgórski. 2001. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Birkhäuser, Boston.
- Marks, R.J., G.L. Wise, D.G. Haldeman and J.L. Whited. 1978. "Detection in Laplace noise." *IEEE Trans. Aerospace Electron. Systems*. 14(6), 866–871.
- Miller, J.H. and J.B. Thomas. 1972. "Detectors for discrete-time signals in non-Gaussian noise." *IEEE Trans. Inform. Theory*. 18(2), 241–250.
- Nevzorov, V.B. 2000. *Records. Mathematical Theory*. Fazis, Moscow (in Russian).
- Wilks, S.S. 1959. "Recurrence of extreme observations." *Journal of American Mathematical Society*, 1(1), 106–112.

AUTHOR BIOGRAPHIES

VLADIMIR BENING is Doctor of Science in physics and mathematics; professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences. His email is bening@yandex.ru.

VLADISLAV SAVUSHKIN is PhD student, Dubna State University. His email is savushkinva@mail.ru

EGOR SHUNKOV is PhD student, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University

ALEXANDER ZEIFMAN is Doctor of Science in physics and mathematics; professor, Head of Department of Applied Mathematics, Vologda State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; principal scientist, Institute of Territories Socio-Economic Development, Russian Academy of Sciences. His email is a_zeifman@mail.ru and his personal webpage at <http://uni-vologda.ac.ru/zai/eng.html>.

VICTOR KOROLEV is Doctor of Science in physics and mathematics, professor, Head of Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

VARIANCE-MEAN MIXTURES AS ASYMPTOTIC APPROXIMATIONS

Victor Korolev
Vladimir Bening
Lomonosov Moscow State University
Leninskie Gory, Moscow, Russia
Institute of Informatics Problems,
Russian Academy of Sciences
Email: victoryukorolev@yandex.ru

Maria Grigoryeva
Parexel International
Osenny Bulvar, 23, Moscow, Russia
Email: maria-grigoryeva@yandex.ru

Andrey Gorshenin
Institute of Informatics Problems,
Russian Academy of Sciences
Vavilova str., 44/2, Moscow, Russia
MIREA, Faculty of Information Technology
Email: agorshenin@ipiran.ru

Alexander Zeifman
Vologda State University
S.Orlova, 6, Vologda, Russia
Institute of Informatics Problems,
Russian Academy of Sciences
Vavilova str., 44/2, Moscow, Russia
Institute of Socio-Economic Development of Territories,
Russian Academy of Sciences
Email: a_zeifman@mail.ru

KEYWORDS

Random sequence, Random index, Transfer theorem, Random sum, Random Lindeberg condition, Statistic constructed from sample with random size, Normal variance-mean mixture.

ABSTRACT

We present a general transfer theorem for random sequences with independent random indexes in the double array limit setting. We also prove its partial inverse providing necessary and sufficient conditions for the convergence of randomly indexed random sequences. Special attention is paid to the cases of random sums of independent not necessarily identically distributed random variables and statistics constructed from samples with random sizes. Using simple moment-type conditions we prove the theorem on convergence of the distributions of such sums to normal variance-mean mixtures.

INTRODUCTION

Random sequences with independent random indexes play an important role in modeling real processes in many fields. Most popular examples of the application of these models usually deal with insurance and reliability theory (Kalashnikov, 1997; Bening and Korolev, 2002), financial mathematics and queuing theory (Bening and Korolev, 2002; Gnedenko and Korolev, 1996), chaotic processes in plasma physics (Korolev and Skvortsova, 2006) where random sums are principal mathematical models. More general randomly indexed random

sequences arrive in the statistics of samples with random sizes. Indeed, very often the data to be analyzed is collected or registered during a certain period of time and the flow of informative events each of which brings a next observation forms a random point process, so that the number of available observations is unknown till the end of the process of their registration and also must be treated as a (random) observation.

The presence of random indexes usually leads to that the limit distributions for the corresponding randomly indexed random sequences are heavy-tailed even in the situations where the distributions of non-randomly indexed random sequences are asymptotically normal see, e. g., (Gnedenko and Korolev, 1996; Bening and Korolev, 2005).

The literature on random sequences with random indexes is extensive, see, e. g., the references above and the references therein. The mathematical theory of random sequences with random indexes and, in particular, random sums, is well-developed. However, there still remain some unsolved problems. For example, necessary and sufficient conditions for the convergence of the distributions of random sums to normal variance-mean mixtures were found only recently for the case of identically distributed summands (see (Korolev, 2013; Grigoryeva and Korolev, 2013)). The case of random sums of *non-identically* distributed random summands and, moreover, more general statistics constructed from samples with random sizes has not been considered yet. At the same time, normal variance-mean mixtures are widely used as mathematical models of statistical regularities in many fields. In particular, in 1977–78 O. Barndorff-Nielsen (Barndorff-Nielsen,

1977), (Barndorff-Nielsen, 1978) introduced the class of *generalized hyperbolic distributions* as a class of special variance-mean mixtures of normal laws in which the mixing is carried out in one parameter since location and scale parameters of the mixed normal distribution are directly linked. The range of applications of generalized hyperbolic distributions varies from the theory of turbulence or particle size description to financial mathematics, see (Barndorff-Nielsen et al., 1982).

The paper is organized as follows. Basic notation is introduced in section "Notation. Auxiliary results". Here an auxiliary result on the asymptotic rapprochement of the distributions of randomly indexed random sequences with special scale-location mixtures is presented. In section "General transfer theorem and its inversion. The structure of limit laws" we present and discuss an improved version of a general transfer theorem for random sequences with independent random indexes in the double array limit setting. We also present its partial inverse providing necessary and sufficient conditions for the convergence of randomly indexed random indexes. Following the lines of (Korolev, 1993), we first formulate a general result improving some results of (Korolev, 1993; Bening and Korolev, 2002) by removing some superfluous assumptions and relaxing some conditions. Special attention is paid to the case where the elements of the basic double array are formed as cumulative sums of independent not necessarily identically distributed random variables. This case is considered in section "Limit theorems for random sums of independent random variables". To prove our results, we use simply tractable moment-type conditions which can be easily interpreted unlike general conditions providing the weak convergence of random sums of non-identically distributed summands in (Szasz, 1972; Kruglov and Korolev, 1990) and (Kruglov and ZhangBo, 2002). In section "A version of the central limit theorem for random sums with a normal variance-mean mixture as the limiting law" we present the theorem on convergence of the distributions of such sums to normal variance-mean mixtures. As a simple corollary of this result we can obtain some results of the recent paper by A.A. Toda "Weak limit of the geometric sum of independent but not identically distributed random variables" (arXiv:1111.1786v2. 2011). That paper demonstrates that there is still a strong interest to geometric sums of non-identically distributed summands and to the application of the skew Laplace distribution which is a normal variance-mean mixture under exponential mixing distribution (Korolev and Sokolov, 2012). Special attention is paid to the case where the elements of the basic double array are formed as statistics constructed from samples with random sizes. This case is considered in section "Limit theorems for statistics constructed from

samples with random sizes".

Unfortunately, to fit the requirements to the size of the paper we had to omit the details of the proofs which will be published elsewhere.

NOTATION. AUXILIARY RESULTS

Assume that all the random variables considered in this paper are defined on one and the same probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. In what follows the symbols $\stackrel{d}{=}$ and \implies will denote coincidence of distributions and weak convergence (convergence in distribution). A family $\{X_j\}_{j \in \mathbb{N}}$ of random variables is said to be *weakly relatively compact*, if each sequence of its elements contains a weakly convergent subsequence. In the finite-dimensional case the weak relative compactness of a family $\{X_j\}_{j \in \mathbb{N}}$ is equivalent to its *tightness*: $\lim_{R \rightarrow \infty} \sup_{n \in \mathbb{N}} \mathbb{P}(|X_n| > R) = 0$ (see, e. g., (Loeve, 1977)).

Let $\{S_{n,k}\}$, $n, k \in \mathbb{N}$, be a double array of random variables. For $n, k \in \mathbb{N}$ let $a_{n,k}$ and $b_{n,k}$ be real numbers such that $b_{n,k} > 0$. The purpose of the constants $a_{n,k}$ and $b_{n,k}$ is to provide weak relative compactness of the family of the random variables $\{Y_{n,k} \equiv (S_{n,k} - a_{n,k})/b_{n,k}\}_{n,k \in \mathbb{N}}$ in the cases where it is required.

Consider a family $\{N_n\}_{n \in \mathbb{N}}$ of nonnegative integer random variables such that for each $n, k \in \mathbb{N}$ the random variables N_n and $S_{n,k}$ are independent. Especially note that we do not assume the row-wise independence of $\{S_{n,k}\}_{k \geq 1}$. Let c_n and d_n be real numbers, $n \in \mathbb{N}$, such that $d_n > 0$. Our aim is to study the asymptotic behavior of the random variables $Z_n \equiv (S_{n,N_n} - c_n)/d_n$ as $n \rightarrow \infty$ and find rather simple conditions under which the limit laws for Z_n have the form of normal variance-mean mixtures. In order to do so we first formulate a somewhat more general result following the lines of (Korolev, 1993), removing superfluous assumptions, relaxing the conditions and generalizing some of the results of that paper.

The characteristic functions of the random variables $Y_{n,k}$ and Z_n will be denoted $h_{n,k}(t)$ and $f_n(t)$, respectively, $t \in \mathbb{R}$.

Let Y be a random variable whose distribution function and characteristic function will be denoted $H(x)$ and $h(t)$, respectively, $x, t \in \mathbb{R}$. Introduce the random variables $U_n = b_{n,N_n}/d_n$, $V_n = (a_{n,N_n} - c_n)/d_n$. Introduce the function

$$g_n(t) \equiv \mathbb{E}h(tU_n) \exp\{itV_n\} = \sum_{k=1}^{\infty} \exp\left\{it \frac{a_{n,k} - c_n}{d_n}\right\} h\left(\frac{tb_{n,k}}{d_n}\right), \quad t \in \mathbb{R}.$$

It can be easily seen that $g_n(t)$ is the characteristic function of the random variable $Y \cdot U_n + V_n$ where the random variable Y is independent of the pair (U_n, V_n) . Therefore, the distribution function

$G_n(x)$ corresponding to the characteristic function $g_n(t)$ is the scale-location mixture of the distribution function $H(x)$:

$$G_n(x) = \mathbf{E}H((x - V_n)/U_n), \quad x \in \mathbb{R}. \quad (1)$$

In the double-array limit setting considered in this paper, to obtain non-trivial limit laws for Z_n we require the following additional *coherency condition*: for any $T \in (0, \infty)$

$$\lim_{n \rightarrow \infty} \mathbf{E} \sup_{|t| \leq T} |h_{n, N_n}(t) - h(t)| = 0. \quad (2)$$

To clarify the sense of the coherency condition, note that if we had usual row-wise convergence of $Y_{n,k}$ to Y , then for any $n \in \mathbb{N}$ and $T \in [0, \infty)$

$$\lim_{k \rightarrow \infty} \sup_{|t| \leq T} |h_{n,k}(t) - h(t)| = 0. \quad (3)$$

So we can say that coherency condition (2) means that "pure" row-wise convergence (3) takes place "on the average" so that that the "row-wise" convergence as $k \rightarrow \infty$ is somehow coherent with the "principal" convergence as $n \rightarrow \infty$.

REMARK 1. It can be easily verified that, since the values under the expectation sign in (2) are nonnegative and bounded (by two), then coherency condition (2) is equivalent to that $\sup_{|t| \leq T} |h_{n, N_n}(t) - h(t)| \rightarrow 0$ in probability as $n \rightarrow \infty$.

LEMMA 1. *Let the family of random variables $\{U_n\}_{n \in \mathbb{N}}$ be weakly relatively compact. Assume that coherency condition (2) holds. Then for any $t \in \mathbb{R}$ we have*

$$\lim_{n \rightarrow \infty} |f_n(t) - g_n(t)| = 0.$$

Lemma 1 makes it possible to use the distribution function $G_n(x)$ (see (1)) as an *accompanying asymptotic* approximation to $F_n(x) \equiv \mathbf{P}(Z_n < x)$. In order to obtain a *limit* approximation, in the next section we formulate and prove the transfer theorem.

GENERAL TRANSFER THEOREM AND ITS INVERSION. THE STRUCTURE OF LIMIT LAWS

THEOREM 1. *Assume that coherency condition (2) holds. If there exist random variables U and V such that the joint distributions of the pairs (U_n, V_n) converge to that of the pair (U, V) :*

$$(U_n, V_n) \Longrightarrow (U, V) \quad (n \rightarrow \infty), \quad (4)$$

then

$$Z_n \Longrightarrow Z \stackrel{d}{=} Y \cdot U + V \quad (n \rightarrow \infty). \quad (5)$$

where the random variable Y is independent of the pair (U, V) .

It is easy to see that relation (5) is equivalent to the following relation between the distribution functions $F(x)$ and $H(x)$ of the random variables Z and Y :

$$F(x) = \mathbf{E}H((x - V)/U), \quad x \in \mathbb{R}, \quad (6)$$

that is, the limit law for normalized randomly indexed random variables Z_n is a scale-location mixture of the distributions which are limiting for normalized non-randomly indexed random variables $Y_{n,k}$. Among all scale-location mixtures, *variance-mean mixtures* attract a special interest (to be more precise, we should speak of *normal variance-mean mixtures*). Let us see how these mixture can appear in the double-array setting under consideration.

Assume that the centering constants $a_{n,k}$ and c_n are in some sense proportional to the scaling constants $b_{n,k}$ and d_n . Namely, assume that there exist $\rho > 0$, $\alpha_n \in \mathbb{R}$ and $\beta_n \in \mathbb{R}$ such that for all $n, k \in \mathbb{N}$ we have

$$a_{n,k} = \frac{b_{n,k}^{\rho+1} \alpha_n}{d_n^\rho}, \quad c_n = d_n \beta_n,$$

and there exist finite limits

$$\alpha = \lim_{n \rightarrow \infty} \alpha_n, \quad \beta = \lim_{n \rightarrow \infty} \beta_n.$$

Then under condition (4) and $n \rightarrow \infty$

$$\begin{aligned} (U_n, V_n) &= \left(\frac{b_{n, N_n}}{d_n}, \frac{a_{n, N_n} - c_n}{d_n} \right) = \\ &= (U_n, \alpha_n U_n^{\rho+1} + \beta_n) \Longrightarrow (U, \alpha U^{\rho+1} + \beta), \end{aligned}$$

so that in accordance with theorem 2 the limit law for Z_n takes the form

$$\mathbf{P}(Z < x) = \mathbf{E}H\left(\frac{x - \beta - \alpha U^{\rho+1}}{U}\right), \quad x \in \mathbb{R}.$$

If $\rho = 1$, then we obtain the "pure" variance-mean mixture

$$\mathbf{P}(Z < x) = \mathbf{E}H\left(\frac{x - \beta - \alpha U^2}{U}\right), \quad x \in \mathbb{R}.$$

We will return to the discussion of convergence of randomly indexed sequences, more precisely, of random sums, to normal scale-location mixtures in Sect. 5.

In order to present the result that is a partial inversion of theorem 1, for fixed random variables Z and Y with the characteristic functions $f(t)$ and $h(t)$ introduce the set $\mathcal{W}(Z|Y)$ containing all pairs of random variables (U, V) such that the characteristic function $f(t)$ can be represented as

$$f(t) = \mathbf{E}h(tU)e^{itV}, \quad t \in \mathbb{R}, \quad (7)$$

and $\mathbf{P}(U \geq 0) = 1$. Whatever random variables Z and Y are, the set $\mathcal{W}(Z|Y)$ is always nonempty since it trivially contains the pair $(0, Z)$. It is easy to see

that representation (7) is equivalent to relation (6) between the distribution functions $F(x)$ and $H(x)$ of the random variables Z and Y .

The set $\mathcal{W}(Z|Y)$ may contain more than one element. For example, if Y is the standard normal random variable and $Z \stackrel{d}{=} W_1 - W_2$ where W_1 and W_2 are independent random variables with the same standard exponential distribution, then along with the pair $(0, W_1 - W_2)$ the set $\mathcal{W}(Z|Y)$ contains the pair $(\sqrt{W_1}, 0)$. In this case $F(x)$ is the symmetric Laplace distribution.

Let $L_1(X_1, X_2)$ be the Lévy distance between the distributions of random variables X_1 and X_2 : if $F_1(x)$ and $F_2(x)$ are the distribution functions of X_1 and X_2 , respectively, then

$$L_1(X_1, X_2) = \inf\{y \geq 0 : F_2(x - y) - y \leq F_1(x) \leq F_2(x + y) + y, \forall x \in \mathbb{R}\}.$$

As is well known, the Lévy distance metrizes weak convergence. Let $L_2((X_1, X_2), (Y_1, Y_2))$ be any probability metric which metrizes weak convergence in the space of two-dimensional random vectors. An example of such a metric is the Lévy-Prokhorov metric (see, e. g., (Zolotarev, 1997)).

THEOREM 2. *Let the family of random variables $\{U_n\}_{n \in \mathbb{N}}$ be weakly relatively compact. Assume that coherency condition (2) holds. Then a random variable Z such that*

$$Z_n \implies Z \quad (n \rightarrow \infty) \quad (8)$$

with some $c_n \in \mathbb{R}$ exists if and only if there exists a weakly relatively compact sequence of pairs $(U'_n, V'_n) \in \mathcal{W}(Z|Y)$, $n \in \mathbb{N}$, such that

$$\lim_{n \rightarrow \infty} L_2((U_n, V_n), (U'_n, V'_n)) = 0. \quad (9)$$

REMARK 2. It should be noted that in (Korolev, 1993) and some subsequent papers a stronger and less convenient version of the coherency condition was used. Furthermore, in (Korolev, 1993) and the subsequent papers the statements analogous to lemma 1 and theorems 1 and 2 were proved under the additional assumption of the weak relative compactness of the family $\{Y_{n,k}\}_{n,k \in \mathbb{N}}$.

LIMIT THEOREMS FOR RANDOM SUMS OF INDEPENDENT RANDOM VARIABLES

Let $\{X_{n,j}\}_{j \geq 1}$, $n \in \mathbb{N}$, be a double array of row-wise independent not necessarily identically distributed random variables. For $n, k \in \mathbb{N}$ denote

$$S_{n,k} = X_{n,1} + \dots + X_{n,k}. \quad (10)$$

If $S_{n,k}$ is a sum of independent random variables, then the condition of weak relative compactness of

the sequence $\{U_n\}_{n \in \mathbb{N}}$ used in the preceding section can be replaced by the condition of weak relative compactness of the family $\{Y_{n,k}\}_{n,k \in \mathbb{N}}$ which is in fact considerably less restrictive. Indeed, let, for example, the random variables $S_{n,k}$ possess moments of some order $\delta > 0$. Then, if we choose $b_{n,k} = (E|S_{n,k} - a_{n,k}|^\delta)^{1/\delta}$, then by the Markov inequality

$$\lim_{R \rightarrow \infty} \sup_{n,k \in \mathbb{N}} P(|Y_{n,k}| > R) \leq \lim_{R \rightarrow \infty} \frac{1}{R^\delta} = 0,$$

that is, the family $\{Y_{n,k}\}_{n,k \in \mathbb{N}}$ is weakly relatively compact.

THEOREM 3. *Assume that the random variables $S_{n,k}$ have the form (10). Let the family of random variables $\{Y_{n,k}\}_{n,k \in \mathbb{N}}$ be weakly relatively compact. Assume that coherency condition (2) holds. Then convergence (8) of normalized random sums Z_n to some random variable Z takes place with some $c_n \in \mathbb{R}$ if and only if there exists a weakly relatively compact sequence of pairs $(U'_n, V'_n) \in \mathcal{W}(Z|Y)$, $n \in \mathbb{N}$, such that condition (9) holds.*

A VERSION OF THE CENTRAL LIMIT THEOREM FOR RANDOM SUMS WITH A NORMAL VARIANCE-MEAN MIXTURE AS THE LIMITING LAW

Let $\{X_{n,j}\}_{j \geq 1}$, $n \in \mathbb{N}$, be a double array of row-wise independent not necessarily identically distributed random variables. As in the preceding section, let $S_{n,k} = X_{n,1} + \dots + X_{n,k}$, $n, k \in \mathbb{N}$. The distribution function of the random variable $X_{n,j}$ will be denoted $F_{n,j}(x)$. Denote $\mu_{n,j} = EX_{n,j}$, $\sigma_{n,j}^2 = DX_{n,j}$ and assume that $0 < \sigma_{n,j}^2 < \infty$, $n, j \in \mathbb{N}$. Denote

$$A_{n,k} = \mu_{n,1} + \dots + \mu_{n,k} \quad (= ES_{n,k}),$$

$$B_{n,k}^2 = \sigma_{n,1}^2 + \dots + \sigma_{n,k}^2 \quad (= DS_{n,k})$$

It is easy to make sure that $ES_{n,N_n} = EA_{n,N_n}$, $DS_{n,N_n} = EB_{n,N_n}^2 + DA_{n,N_n}$, $n \in \mathbb{N}$. In order to formulate a version of the central limit theorem for random sums with the limiting distribution being a normal variance-mean mixture, as usual, are centered by their expectations and normalized by their mean square deviations and put $a_{n,k} = A_{n,k}$, $b_{n,k} = \sqrt{B_{n,k}^2}$, $n, k \in \mathbb{N}$. Although it would have been quite natural to normalize random sums by their mean square deviations as well, for simplicity we will use slightly different normalizing constants and put $d_n = \sqrt{EB_{n,N_n}^2}$.

Let $\Phi(x)$ be the standard normal distribution function,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-z^2/2} dz, \quad x \in \mathbb{R}.$$

THEOREM 4. Assume that the following conditions hold:

(i) for every $n \in \mathbb{N}$ there exist real numbers α_n such that

$$\mu_{n,j} = \frac{\alpha_n \sigma_{n,j}^2}{\sqrt{\mathbb{E}B_{n,N_n}^2}}, \quad n, j \in \mathbb{N},$$

and

$$\lim_{n \rightarrow \infty} \alpha_n = \alpha, \quad 0 < |\alpha| < \infty;$$

(ii) (the random Lindeberg condition) for any $\epsilon > 0$

$$\mathbb{E} \left[\frac{1}{B_{n,N_n}^2} \sum_{j=1}^{N_n} \int_{|x - \mu_{n,j}| > \epsilon B_{n,N_n}} (x - \mu_{n,j})^2 dF_{n,j}(x) \right] \rightarrow 0$$

as $n \rightarrow \infty$.

Then the convergence of the normalized random sums

$$\frac{S_{n,N_n}}{\sqrt{\mathbb{E}B_{n,N_n}^2}} \Longrightarrow Z \quad (11)$$

to some random variable Z as $n \rightarrow \infty$ takes place if and only if there exists a random variable U such that

$$\mathbb{P}(Z < x) = \mathbb{E}\Phi\left(\frac{x - \alpha U}{\sqrt{U}}\right), \quad x \in \mathbb{R}, \quad (12)$$

and

$$\frac{B_{n,N_n}^2}{\mathbb{E}B_{n,N_n}^2} \Longrightarrow U \quad (n \rightarrow \infty). \quad (13)$$

The proof uses a result of V. V. Petrov (Petrov, 1979) improved in (Korolev and Popov, 2012), the reasoning used to prove theorem 7 in Sect. 3, Chapt. V of (Petrov, 1987) and relation (3.8) and the fact recently proved in (Korolev, 2013) that normal variance-mean mixtures are identifiable.

REMARK 3. In accordance with what has been said in remark 1, the random Lindeberg condition (ii) can be used in the following form: for any $\epsilon > 0$

$$\frac{1}{B_{n,N_n}^2} \sum_{j=1}^{N_n} \int_{|x - \mu_{n,j}| > \epsilon B_{n,N_n}} (x - \mu_{n,j})^2 dF_{n,j}(x) \rightarrow 0$$

in probability as $n \rightarrow \infty$.

LIMIT THEOREMS FOR STATISTICS CONSTRUCTED FROM SAMPLES WITH RANDOM SIZES

Let $\{X_{n,j}\}_{j \geq 1}$, $n \in \mathbb{N}$, be a double array of row-wise independent not necessarily identically distributed random variables. For $n, k \in \mathbb{N}$ let $T_{n,k} = T_{n,k}(X_{n,1}, \dots, X_{n,k})$ be a statistic, i.e., a real-valued measurable function of $X_{n,1}, \dots, X_{n,k}$. For each $n \geq 1$ we define a r.v. T_{n,N_n} by setting $T_{n,N_n}(\omega) \equiv T_{n,N_n(\omega)}(X_{n,1}(\omega), \dots, X_{n,N_n(\omega)}(\omega))$, $\omega \in \Omega$.

Let θ_n be real numbers, $n \in \mathbb{N}$. In this section we will assume that the random variables $S_{n,k}$ have the form $S_{n,k} = T_{n,k} - \theta_n$, $n, k \in \mathbb{N}$. Concerning the normalizing constants we will assume that there exist finite real numbers $\alpha, \alpha_n, \beta, \beta_n, \sigma_n > 0$ such that

$$\alpha_n \rightarrow \alpha, \quad \beta_n \rightarrow \beta \quad (n \rightarrow \infty) \quad (14)$$

and for all $n, k \in \mathbb{N}$

$$\begin{aligned} b_{n,k} &= \frac{1}{\sigma_n \sqrt{k}}, & d_n &= \frac{1}{\sigma_n \sqrt{n}}, \\ a_{n,k} &= \frac{\alpha_n \sqrt{n}}{\sigma_n k}, & c_n &= \frac{\beta_n}{\sigma_n \sqrt{n}} \end{aligned} \quad (15)$$

so that

$$Y_{n,k} = \sigma_n \sqrt{k}(T_{n,k} - \theta_n) - \alpha_n \sqrt{n/k}$$

and

$$Z_n = \sigma_n \sqrt{n}(T_{n,N_n} - \theta_n) - \beta_n.$$

As this is so, σ_n^2 can be regarded as the asymptotic variance of $T_{n,k}$ as $k \rightarrow \infty$ whereas the bias of $T_{n,k}$ is $\alpha_n \sqrt{n}/(k\sigma_n)$.

The l_1 -distance between distribution functions P_1 and P_2 will be denoted $\|P_1 - P_2\|$:

$$\|P_1 - P_2\| = \int_{-\infty}^{\infty} |P_1(x) - P_2(x)| dx.$$

Recall that the distribution function of the random variable $Y_{n,k}$ is denoted $H_{n,k}(x)$. Let $\Phi(x)$ be the standard normal distribution function. In what follows we will assume that the statistic $T_{n,k}$ is asymptotically normal in the following sense: for any $\gamma > 0$

$$\lim_{n \rightarrow \infty} \mathbb{E}\|H_{n,N_n} - \Phi\| = 0. \quad (16)$$

THEOREM 5. Let the family of random variables $\{n/N_n\}_{n \in \mathbb{N}}$ be weakly relatively compact, the normalizing constants have the form (15) and satisfy condition (14). Assume that the statistic $T_{n,k}$ is asymptotically normal so that condition (16) holds. Then a random variable Z such that

$$\sigma_n \sqrt{n}(T_{n,N_n} - \theta_n) - \beta_n \Longrightarrow Z \quad (n \rightarrow \infty)$$

exists if and only if there exists a nonnegative random variable W such that ($x \in \mathbb{R}$)

$$\mathbb{P}(Z < x) = \int_0^{\infty} \Phi\left(\frac{x - \beta - \alpha w}{\sqrt{w}}\right) d\mathbb{P}(W < w),$$

and

$$\mathbb{P}(N_n < nx) \Longrightarrow \mathbb{P}(W^{-1} < x) \quad (n \rightarrow \infty).$$

CONCLUDING REMARKS

The class of normal variance-mean mixtures is very wide and, in particular, contains the class of generalized hyperbolic distributions which, in turn, contains (a) symmetric and skew Student distributions (including the Cauchy distribution) with inverse gamma mixing distributions; (b) variance gamma distributions (including symmetric and non-symmetric Laplace distributions) with gamma mixing distributions; (c) normal/inverse Gaussian distributions with inverse Gaussian mixing distributions including symmetric stable laws. By variance-mean mixing many other initially symmetric types represented as pure scale mixtures of normal laws can be skewed, e. g., as it was done to obtain non-symmetric exponential power distributions in (Grigoryeva and Korolev, 2013).

According to theorem 4, all these laws can be limiting for random sums of independent non-identically distributed random variables. For example, to obtain the skew Student distribution for Z it is necessary and sufficient that in (12) and (13) the random variable U has the inverse gamma distribution (Korolev and Sokolov, 2012). To obtain the variance gamma distribution for Z it is necessary and sufficient that in (12) and (13) the random variable U has the gamma distribution (Korolev and Sokolov, 2012). In particular, for Z to have the asymmetric Laplace distribution it is necessary and sufficient that U has the exponential distribution.

REMARK 4. Note that the non-random sums in the coherency condition are centered, whereas in (11) the random sums are not centered, and if $\alpha \neq 0$, then the limit distribution for random sums becomes skew unlike usual non-random summation, where the presence of the systematic bias of the summands results in that the limit distribution becomes just shifted. So, if non-centered random sums are used as models of some real phenomena and the limit variance-mean mixture is skew, then it can be suspected that the summands are actually biased.

REMARK 5. In limit theorems of probability theory and mathematical statistics, centering and normalization of random variables are used to obtain non-trivial asymptotic distributions. It should be especially noted that to obtain reasonable approximation to the distribution of the basic random variables (in our case, S_{n, N_n}), both centering and normalizing values should be non-random. Otherwise the approximate distribution becomes random itself and, say, the problem of evaluation of quantiles becomes senseless.

ACKNOWLEDGEMENTS

The research is supported by the Russian Foundation for Basic Research (projects 12-07-00115a,

12-07-00109a, 14-07-00041a) and the Grant of the President of the Russian Federation (project MK-4103.2014.9).

REFERENCES

- Barndorff-Nielsen, O.E. 1977. "Exponentially decreasing distributions for the logarithm of particle size." *Proc. Roy. Soc. Lond., Ser. A* 353, 401-419.
- Barndorff-Nielsen, O.E. 1978. "Hyperbolic distributions and distributions of hyperbolae." *Scand. J. Statist.* 5, 151-157.
- Barndorff-Nielsen, O.E., J. Kent, M. Sørensen. 1982. "Normal variance-mean mixtures and z -distributions." *Int. Statist. Rev.* 50, No.2, 145-159.
- Bening, V.E.; and V.Yu. Korolev. 2002. *Generalized Poisson Models and their Applications in Insurance and Finance*. VSP, Utrecht.
- Bening, V.E. and V.Yu. Korolev. 2005. "On an application of the Student distribution in the theory of probability and mathematical statistics." *Theory of Probability and its Applications* 49, No.3, 377-391.
- Gnedenko, B.V.; and V.Yu. Korolev. 1996. *Random Summation: Limit Theorems and Applications*. CRC Press, Boca Raton.
- Grigoryeva, M.E. and V.Yu. Korolev. 2013. "On convergence of the distributions of random sums to asymmetric exponential power laws." *Informatics and its Applications* 7, No.4, 66-74.
- Loève, M. 1977. *Probability Theory. 3rd ed.* Springer, New York.
- Kalashnikov, V.V. 1997. *Geometric Sums: Bounds for Rare Events with Applications*. Academic Publishers, Dordrecht, Kluwer.
- Korolev, V.Yu. 1993. "On limit distributions of randomly indexed random sequences". *Theory of Probability and its Applications* 37, No.3, 535-542.
- Korolev, V.Yu. 2013. "Generalized hyperbolic laws as limit distributions for random sums." *Theory of Probability and its Applications* 58, No.1, 117-132.
- Korolev, V.Yu. and S.V. Popov. 2012. "Improvement of convergence rate estimates in the central limit theorem under weakened moment conditions." *Doklady Mathematics* 86, No.1, 506-511.
- Korolev, V.Yu.; and N.N. Skvortsova (Eds). 2006. *Stochastic Models of Structural Plasma Turbulence*. VSP, Utrecht.

Korolev, V.Yu. and I.A. Sokolov. 2012. "Skew Student distributions, variance gamma distributions and their generalizations as asymptotic approximations." *Informatics and its Applications* 6, No.1, 2-10.

Kruglov, V.M.; and V.Yu. Korolev. 1990. *Limit Theorems for Random Sums*. Moscow University Publishing House, Moscow.

Kruglov, V.M. and Zhang Bo. 2002. "Weak convergence of random sums." *Theory of Probability and its Applications* 46, No.1, 39-57.

Petrov, V.V. 1979. "A limit theorem for sums of independent, nonidentically distributed random variables." *Journal of Soviet Mathematics* 20, No.3, 2232-2235.

Petrov, V.V. 1987. *Limit Theorems for Sums of Independent Random Variables*. Nauka, Moscow.

Szász, D. 1972. "Limit theorems for the distributions of the sums of a random number of random variables." *Annals of Mathematical Statistics* 43, No.6, 1902-1913.

Zolotarev, V.M.; and N.N. Skvortsova. 1997. *Modern Theory of Summation of Random Variables*. VSP, Utrecht.

is: maria-grigoryeva@yandex.ru

ALEXANDER ZEIFMAN is Doctor of Science in physics and mathematics; professor, Dean of the Faculty of Applied Mathematics and Computer Technologies, Vologda State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; leading scientist, Institute of Socio-Economic Development of Territories, Russian Academy of Sciences. His e-mail address is: a.zeifman@mail.ru

AUTHOR BIOGRAPHIES

VICTOR KOROLEV is Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; leading scientist, Institute of Informatics Problems, Russian Academy of Sciences. His e-mail address is: victoryukorolev@yandex.ru

VLADIMIR BENING is Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences. His e-mail address is: bening@yandex.ru

ANDREY GORSHENIN is Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences; associate professor, Faculty of Information Technology, Moscow State Institute of Radio Engineering, Electronics and Automation. His e-mail address is: agorshenin@ipiran.ru

MARIA GRIGORYEVA is biostatistician II, Parexel International, Moscow. Her e-mail address

ANALYTICAL MODELLING AND SIMULATION FOR PERFORMANCE EVALUATION OF SIP SERVER WITH HYSTERETIC OVERLOAD CONTROL

Konstantin E. Samouylov
Pavel O. Abaev (speaker)
Yuliya V. Gaidamaka
Telecommunication Systems Department
Peoples' Friendship University of Russia
Miklukho-Maklaya str., 6,
117198, Moscow, Russia
Email: {ksam, pabaev, ygaidamaka}@sci.pfu.edu.ru

Alexander V. Pechinkin
Rostislav V. Razumchik
Institute of Informatics Problems of RAS
Vavilova, 44-1,
119333, Moscow, Russia
Email: apechinkin@ipiran.ru,
rrazumchik@iee.org

KEYWORDS

Signalling network, SIP, hop-by-hop overload control, threshold, hysteretic load control, queuing model.

ABSTRACT

Major standards organizations, ITU, ETSI, and 3GPP have all adopted SIP as a basic signalling protocol for NGN. The current SIP overload control mechanism is unable to prevent congestion collapse and may spread the overload condition throughout the network. In this paper, we investigate one of the implementations of loss based overload control scheme developed by IETF work group which uses hysteretic load control technique on the server side for preventing its overloading. Two different approaches to calculate performance measure of SIP server are introduced. We follow an analytical modelling approach to construct and analyse SIP server model in the form of queuing system with finite buffer occupancy and two-level hysteretic overload control. The formulas for stationary probabilities and the mean return time in the set of normal states were obtained. Simulation is the other approach which allows to eliminate disadvantages of analytical modelling. At present, there is no simulator for modelling of SIP servers in overload conditions with an application of overload control mechanisms which are currently under development by IETF. Approaches to its programming implementations which reflects the protocols and functions that are fully or partially built into the original SIP systems are proposed in the paper.

INTRODUCTION

SIP is an application-layer signaling protocol for creating, modifying, and terminating sessions with one or more participants. In November 2000, SIP was accepted as a 3GPP signaling protocol and main protocol of the IMS architecture. In 2002, recommendation (RFC 3261, 2002) which determines the current protocol form was accepted. The rapid development of the market for services based on the SIP protocol and the growing user needs have revealed a number of shortcomings in the protocol, specifically, in the base overload control mecha-

nism (mechanism 503). In 2009, Rosenberg, one of the protocol designers, demonstrated in (RFC 5390, 2008) the protocols main shortcomings in regard to overload prevention and formulated the main requirements toward the future overload control mechanism. In mid-2010, the SOC working group was created within the IETF Committee. Its work aims at creating overload prevention mechanisms. The first result of their work was the document (IETF draft SIP Rate Control, 2012) permanently accepted in August 2011. The document provides a discussion of the available types of overload control mechanisms local, hop-by-hop, and end-to-end, a classification of SIP networks, and presents the overall architecture of overload-control systems. The SOC group's work focuses on developing two hop-by-hop schemes for overload control as this type of mechanisms has a number of indisputable advantages over the other two types (IETF draft SIP Overload Control, 2013; IETF draft SIP Rate Overload Control, 2013). At present, two overload control schemes have been proposed one with flow sifting on the sender side (LBOC, Loss-based overload control) and one with restricting the flow rate of signaling messages (RBOC, Rate-based overload control). However, only the basic principles were described in SOCs' documents and methods for calculation of the control parameters were not specified. The control parameters can be determined based on analysis of mathematical models or as the results of simulation modeling. As the processes going on in the SIP networks are difficult to describe mathematically and they depend on a large number of different factors, the task needs to be solved through the creation of a simulator.

This paper is organized as follows. In Section 1 we recall SIP client-server model, list of overload control problems that arise because of defect in built-in overload control mechanism, solution requirements that address the problems, and define quality of signaling metrics according to recent standards that fulfill requirements. In Section 2 the hysteretic control mechanism for hop-by-hop overload control based on LBOC scheme is described. In Section 3 we build and analyze a queuing model with threshold control using the embedded

Markov chain apparatus. Finally in Section 4 the design and architecture of SIP network simulator is described.

SIP OVERVIEW

SIP is a request/response-based protocol. End users are represented by user agents (UAs), which take the role of a user agent client (UAC) or user agent server (UAS) for a request/response pair. A UAC creates a SIP request and sends it to a UAS. On its way, a SIP request typically traverses one or more SIP proxy servers. The main purpose of a SIP server is to route a request one hop closer to its destination. Responses trace back the path the request has taken.

The UAC sends an INVITE request to the UAS to initiate a SIP session. Each server on the path confirms the reception of this request by returning a 100 Trying response to the previous hop. Instead of forwarding a request, a SIP server can reject it if it is unwilling or unable to forward the request. Once the request is received by the UAS, it typically responds with a 180 Ringing response to indicate that the called user is being alerted and a 200 OK response when the user has accepted the session. After the 200 OK is received by the UAC, it sends an ACK request to complete the three way handshake of an INVITE transaction. The INVITE request is the only SIP request that uses a three way handshake. Sessions can be terminated at any time by sending a BYE request, which is confirmed with a 200 OK response.

SIP Overload Problem and Servers' KPIs

With the increasing deployment of SIP based systems and therefore an increasing number of users that utilize new services, the chance for overload on some nodes rises. A SIP server overload occurs if a SIP server does not have sufficient resources to process all incoming SIP messages. Several reasons, including Poor Capacity Planning, Component Failures, Avalanche Restart or Flash Crowds and the list of the problems that arise as a result of the 503 mechanism are presented in (RFC 5390, 2008), including load amplification problem, underutilization or the Off/On retry-after problem.

Rosenberg formulated 23 requirements to overload control mechanisms in (RFC 5390, 2008); mechanisms matching them will be able to predict and to avoid or quickly to cope with an overload on the server.

It is therefore important to continuously track the current load on all SIP nodes of a telecommunication operator to be able to detect possibly dangerous situations on the one hand as well as to apply load reducing procedures if necessary on the other hand. For this purpose, continuous measurements on respective hosts are required and thus, measurement metrics need to be defined. The IETF has created a working group called IP Performance Metrics (IPPM) and developed a set of metrics that can be applied to the quality, performance and reliability of Internet data delivery services. The base framework for these metrics is defined in RFC 2330 (RFC 2330, 1998). How-

ever, this framework is designed for the network layer and is not suitable for application layer protocols. With further progress, the IETF has started another working group in order to define metrics for the remaining layers and named it Performance Metrics for Other Layers (PMOL). This group has standardized its first approach within (RFC 6067, 2010), but, however, the scope of this document is limited to an end-to-end perspective. This perspective does not allow to profile the performance of intermediate entities in the signaling path because it provides only an outside view. Still, it is necessary to define the used measurement metric between two hops in order to detect possible performance problems on a specific intermediate hop. Additionally it is required that the analysis of measurement results based upon these metrics permits a clear differentiation between a overloaded and a non-overloaded system.

The Quality of Signaling metrics, defined in (Happenhofer, 2010) and (Happenhofer2, 2010) have been chosen because they fulfill the mentioned above requirements. The metrics are essential for the following performance analysis, that are:

- Final Response Delay (fRpD) – Time span between sending a request until a final response is received;
- Request Transmits – Total number of requests sent per transaction;
- Success Rate – Ratio of number of successful transactions to injected transactions.

CONCEPT OF HOP-BY-HOP OVERLOAD CONTROL

Current work of the SOC group is focused on the development of two hop-by-hop overload control schemes – Loss-based overload control and Rate-based overload control.

The basic idea of LBOC scheme is that the sending entity (SE) reduces the number of messages on RE's request which will be send to the receiving entity (RE) by specified in the request amount of the total number of messages. RBOC scheme operates in the following way: RE informs SE about the maximum message rate which RE would like to receive from SE within a specified period of time. RE sends the control information to SE periodically depending on RE load changes. Both of these schemes based on the idea of feedback control loop between all neighbouring SIP servers that directly exchange traffic. Each loop controls only two entities. The Actuator is located on the sending entity and throttles the traffic if necessary. The receiving entity has the Monitor which measures the current server load.

The four Via header parameters ('oc', 'oc-algo', 'oc-validity' and 'oc-seq') are introduced in (IETF draft SIP Overload Control, 2013) to transfer the control information between two adjacent entities. The integer parameter 'oc' consisting of 10 digits and its value defines what percentage of the total number of SIP requests are subject to reduction at the SE when the loss-based scheme is used.

Analogously, when the rate-based scheme is used it indicates that the client should send SIP requests at a rate of ‘oc’-value SIP requests or fewer per second. ‘oc-algo’ parameter defines the scope of algorithms supported by SE, e.g. oc-algo=“loss”, “rate”. ‘oc-validity’ parameter contains a value that indicates an interval of time (measured in milliseconds) that the load reduction specified in the value of the ‘oc’ parameter should be in effect, its default value is 500 ms. ‘oc-sequence’ is the sequence number associated with the ‘oc’ parameter, timestamps usually use as its value.

Threshold Overload Control on the Server Side

As a criteria of determining the choice of moments for sending messages with control information from SE to RE we propose to use hysteretic control technique. The system during operation changes its state depending on the total number of messages n present in it. Choose arbitrary numbers L and H such that $0 < L < H < R$, where R is the buffer capacity. When the system starts to work it is empty, ($n = 0$), and as long as the total number of messages in the system remains below $H - 1$, system is considered to be in normal state, ($s = 0$). When total number of messages exceeds $H - 1$ for the first time, the system changes its state to overload, ($s = 1$), and RE informs SE that traffic load should be reduced: it stays in it as long as the number of messages remains between L and $R - 1$. Being in overload state, RE’s system waits till the number of messages drops down below L after which it changes its state back to normal and informs SE about changes, or exceeds $R - 1$ after which it changes its state to blocking, ($s = 2$), and ask SE for temporary suspension of sending SIP requests. When the total number of messages drops down below $H + 1$, system’s state changes back to overload, and RE informs SE that the process of sending of messages can be resumed with the current limitations. Input load function $\lambda(s, n)$ is schematically depicted in Fig. 1.

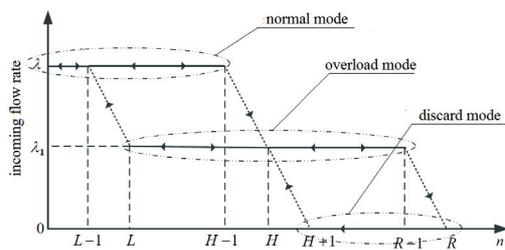


Figure 1: Hysteretic load control

Default Algorithm on the Client Side for LBOC case

In the case of LBOC scheme the default algorithm for throttling incoming to the server traffic is used on the client side. The idea of the algorithm presented in (IETF draft SIP Overload Control, 2013) is to sift the client’s outgoing flow. Let us consider the example of the implementation of the algorithm.

The client maintains two types of requests – the priority and non-priority. Prioritization of messages is done in accordance with local policies applicable to each SIP-server. In situations where the client has to sift the outgoing flow, it first reduces non-priority messages, and then if the buffer contains only priority messages and further reduction is still needed, the client reduces the priority messages.

Under overload condition, the client converts the value of the ‘oc’= q parameter to a value that it applies to non-priority requests. Let N_1 denote the number of priority messages and N_2 denote the number of the non-priority messages in the client’s buffer. The client should reduce the non-priority messages with probability $q_2 = \min \left\{ 1, q \frac{N_1 + N_2}{N_2} \right\}$ and the priority messages with probability $q_1 = \frac{q(N_1 + N_2) - q_2 N_2}{N_1}$ if necessary to get an overall reduction of the ‘oc’ value.

To affect the reduction rate with probability q_2 from the non-priority messages, the client draws a random number between 1 and 100 for the request picked from the first category. If the random number is less than or equal to converted value of the “oc” parameter, the request is not forwarded; otherwise the request is forwarded. Recalculation of probabilities is performed periodically every 5-10 seconds by getting the value of the counters N_1 and N_2 .

QUEUING MODEL WITH THRESHOLDS CONTROL

Consider a single-queue, single-server system with general service time distribution function which is denoted by $B(x)$ and hysteretic load control. Let denote the Laplace-Stieltjes transform (LST) of $B(x)$ by $\beta(s)$ and the mean service time by $b < \infty$.

Two types of customers (say, type 1 and type 2) arrive at the system in batches (each) in accordance with a Poisson process with rates λ_1 and λ_2 respectively. Henceforth the compound flow rate is denoted by $\lambda = \lambda_1 + \lambda_2$. Each batch has a random number of customers and the probability that arriving batch of type k , $k = 1, 2$, customers contains exactly n , $n \geq 1$, customers is $\omega_{k,n}$.

In further analysis it is assumed that thresholds’ values are chosen in such a way that inequalities $H - L \geq 1$ and $R - H \geq 2$ hold.

We turn to the $M^{[X]}|G|1|(L, H)|(H, R)$ queue operating according to the policy that the system may switch between operating modes only at the time instant of a customer departure. If in normal mode just before the customer departure the total number of customers in the system equals H , then the system switches to the overload mode. Similarly if in overload mode just before the customer departure the total number of customers in the system equals H , then the system switches to the discard mode.

Let us denote $X(t)$ a two-dimensional stochastic pro-

cess with set of states

$$S = \left\{ (j, s) \left| \begin{array}{ll} j = 0, \dots, R, & s = 0 \\ j = L, \dots, R, & s = 1 \\ j = H + 1, \dots, R - 1, & s = 2 \end{array} \right. \right\}$$

and its subsets $S_i = \{(j, s) \in S | s = i\}$, $i = 0, 1, 2$, where j is the number of customers in the system and s indicates system operating mode.

Take the service completion epochs to be $0 < t_1 < t_2 < \dots$, where t_n is the instant of the n th customer departure. Then the discrete-time process embedded at customer departure epochs $X(t_n + 0)$ emerges a Markov chain with set of states

$$\tilde{S} = \left\{ (j, s) \left| \begin{array}{ll} j = 0, \dots, H - 2, & s = 0 \\ j = L, \dots, R - 2, & s = 1 \\ j = H + 1, \dots, R - 1, & s = 2 \end{array} \right. \right\};$$

$$\tilde{S}_i = \{(j, s) \in \tilde{S} | s = i\}, \quad i = 0, 1, 2.$$

Let us denote $\{p_{j,s}\}$ and $\{q_{j,s}\}$ stationary distribution of $X(t)$ and $X(t_n + 0)$ respectively:

$$p_{j,s} = \lim_{t \rightarrow \infty} P\{X(t) = (j, s)\}, \quad (j, s) \in S;$$

$$q_{j,s} = \lim_{n \rightarrow \infty} P\{X(t_n + 0) = (j, s)\}, \quad (j, s) \in \tilde{S}.$$

To obtain transition probabilities of the Markov chain we introduce the probability that in operating mode s during the service time of a customer exactly k batches will arrive the system:

$$\beta_k^s = \frac{\lambda_k^s}{k!} \beta^{(k)}(\lambda_s), \quad s = 0, 1, \quad k \geq 0.$$

To express transition probabilities of the Markov chain $X(t_n + 0)$ we introduce the following auxiliary variables:

$$\omega_i^0 = \delta_i, \omega_i^k = \sum_{n=0}^i \omega_{i-n}^{k-1} \frac{\lambda_1 \omega_{1,n} + \lambda_2 \omega_{2,n}}{\lambda}, \quad k \geq 1, i \geq 0,$$

where δ_i is the Kronecker delta ($\delta_i = 1$ if $i = 0$, or 0 otherwise).

Let us denote by α_i^s , $s = 0, 1$, $i \geq 0$, — the probability that in operating mode s exactly i new customers arrive during the time of service of a customer; A_i^s , $s = 0, 1$, $i \geq 0$, — the probability that in operating mode s not less than i new customers arrive during the time of service of a customer; γ_i , $i \geq 0$, — the probability that immediately after the departure of the customer arrived when the system was empty, there will be exactly i customers in the system:

$$\alpha_i^s = \sum_{k=0}^i \beta_k^s \omega_i^k, \quad s = 0, 1, \quad i \geq 0,$$

$$A_i^s = \sum_{k=i}^{\infty} \alpha_k^s, \quad s = 0, 1, \quad i \geq 0,$$

$$\gamma_i = \sum_{k=1}^{i+1} \omega_k^0 \alpha_{i-k+1}^0, \quad i \geq 0.$$

Thus the equilibrium equations for probability distribution $\{q_{j,s}\}$ takes the form of

$$\begin{aligned} q_{j,0} &= q_{0,0} \gamma_j + \sum_{i=1}^{\min(j+1, H-2)} q_{i,0} \alpha_{j-i+1}^0 + \\ &+ \delta_{j-L+1} q_{L,1} \alpha_0^1, \quad j = 0, \dots, H-2, \\ q_{j,1} &= q_{0,0} \gamma_j + \sum_{i=1}^{H-2} q_{i,0} \alpha_{j-i+1}^0 + \\ &+ \sum_{i=L}^{\min(j+1, R-2)} q_{i,1} \alpha_{j-i+1}^1 + \\ &+ \delta_{j-H} q_{H+1,2}, \quad j = H-1, \dots, R-2, \\ q_{j,1} &= \sum_{i=L}^{j+1} q_{i,1} \alpha_{j-i+1}^1, \quad j = L, \dots, H-2, \\ q_{R-1,2} &= q_{0,0} \sum_{i=R-1}^{\infty} \gamma_i + \\ &+ \sum_{i=1}^{H-2} q_{i,0} A_{R-i}^k + \sum_{i=L}^{R-2} q_{i,1} A_{R-i}^1, \\ q_{j,2} &= q_{R-1,2}, \quad j = H+1, \dots, R-2. \end{aligned} \quad (1)$$

The probability $q_{0,0}$ is determined from the normalization condition.

Stationary state distribution

We use the renewal theory to receive the stationary queue length distribution of the corresponding stochastic process from the stationary queue length distribution of the embedded Markov chain.

The stationary mean T of the time interval between neighboring instants t_n and t_{n+1} is defined by the formula

$$T = b + \frac{1}{\lambda_0} q_{0,0}.$$

We also denote $\nu = 1/T$,

$$\tilde{\beta}_k^s = \frac{\lambda_k^s}{k!} \tilde{\beta}^{(k)}(\lambda_s), \quad s = 0, 1, \quad k \geq 0,$$

$$\tilde{\alpha}_i^s = \sum_{k=0}^i \tilde{\beta}_k^s \omega_i^k, \quad s = 0, 1, \quad i \geq 0,$$

$$\tilde{A}_i^s = \sum_{k=i}^{\infty} \tilde{\alpha}_k^s, \quad s = 0, 1, \quad i \geq 0.$$

$$\tilde{\gamma}_i = \sum_{k=1}^i \omega_k^0 \tilde{\alpha}_{i-k}^0, \quad i \geq 1.$$

The following theorem contains formulas for calculating the stationary distribution $\{p_{j,s}\}$.

Theorem. Stationary probabilities of the stochastic process $X(t)$ are given by

$$p_{0,0} = \frac{\nu}{\lambda_0} q_{0,0},$$

$$p_{j,0} = \nu \left(\tilde{\gamma}_j q_{0,0} + \sum_{i=1}^{\min(j,H-2)} \tilde{\alpha}_{j-i}^0 q_{i,0} \right), \quad j = 1, \dots, R-1,$$

$$p_{R,0} = \nu \left(\sum_{i=R}^{\infty} \tilde{\gamma}_i q_{0,0} + \sum_{i=1}^{H-2} \tilde{A}_{j-i}^0 q_{i,0} \right),$$

$$p_{j,1} = \nu \sum_{i=L}^{\min(j,R-2)} \tilde{\alpha}_{j-i}^1 q_{i,1}, \quad j = L, \dots, R-1,$$

$$p_{R,1} = \nu \sum_{i=L}^{R-2} \tilde{A}_{j-i}^1 q_{i,1},$$

$$p_{j,2} = \nu b q_{R-1,2}, \quad j = H+1, \dots, R-1.$$

Performance measures

We denote by $P(S_1) = \sum_{j=L}^R p_{j,1}$ the stationary probability of the system being in overload mode, by $P(S_2) = \sum_{j=H+1}^{R-1} p_{j,2}$ the system being in discard mode, by $P(S_0) = \sum_{j=0}^{H-2} p_{j,0}$ the system being in normal mode.

The mean control cycle time is inverse to stationary intensity of instants of control cycle starts. Since the control cycle starts when the system passes from state $(L, 1)$ to state $(L-1, 0)$, then the stationary intensity of instants of control cycle starts is equal to the stationary intensity of passes from state $(L, 1)$ to state $(L-1, 0)$ is defined as follows:

$$\mu = \nu q_{L,1} \int_0^{\infty} e^{-\lambda_1 x} dB(x) = \nu \beta_0^1 q_{L,1} = \tau^{-1}.$$

Thus mean time τ_{12} the system spends in overload and discard set of states during one control cycle can be calculated by the following formula

$$\tau_{12} = \frac{P(S_1) + P(S_2)}{P(S_0) + P(S_1) + P(S_2)} \cdot \frac{1}{\nu \beta_0^1 q_{L,1}}.$$

Numerical Example

Below we provide a numerical example relating to the mean return time of the system from the overload mode to the normal load mode in the case of small dimension, i.e. with the values of the thresholds $L = 8$, $H = 12$ and $R = 20$. We consider two versions of the service time distribution: an exponential service time and constant service time with the same mean value of $b = 1$. Let the distribution of the number of customers in the batch be $\omega_{k,n} = 0.2$, where $k = 1, 2$ and $n = \overline{1, 5}$. Let the intensity of the input flow be $\lambda = 2/3$, and since the average number of customers in the batch is equal to 3, then the offered load intensity takes a value $\rho = 2$.

Fig. 2 presents the results of the calculations in the form of graphs showing the dependence of mean return time τ_{12} on the dropping probability q for the two policies of overload control. The mean service time is taken per time unit. One can note that the mean return time for an exponential service time is less than those for a constant time for the values of $q < 0.6$

SIP NETWORK SIMULATION

The network testbed that we used to generate calls and collect measurement data consists of UAC, UAS and SIP proxy. The instance of SIPp tool modified to support Loss based overload control scheme was run on UAC and UAS. On the server side the SIP server instance that supported basic FSM for successful call setup scenario was run.

Figure 3 illustrates the steps a packet takes as it moves from the device driver through the Linux kernel to the SIP layer and back to the device driver, and then we introduce the methodology we used to measure all the components of the packet service and waiting times.

Based on this figure, there are three distinct entities involved in processing a SIP packet. The first one of them is Kernel Network Stack, which provides the procedure of receiving of packets. As soon as a packet is received from the physical device, it arrives at the device driver and is transferred to a ring buffer in kernel space. The packet then undergoes processing within the Linux kernel stack before it is handed to the application – SIP layer.

Application Layer provides SIP Packet Processing procedures. The SIP layer has a blocking loop waiting for packets to arrive on the socket. Processing at the SIP layer consists of two parts: one that is common to all messages and one that is message dependant. During the common part, the message is parsed, it is classified as a SIP request or response, and then certain operations are performed. Since the proxy operates in stateful mode, a lookup is performed to determine if the SIP transaction is already present. If not, a new transaction is created, otherwise the existing transaction is returned and the SIP message-specific processing phase begins. Once the forwarding, reply or relay decision has been made, a send buffer is created with the updated data for the SIP

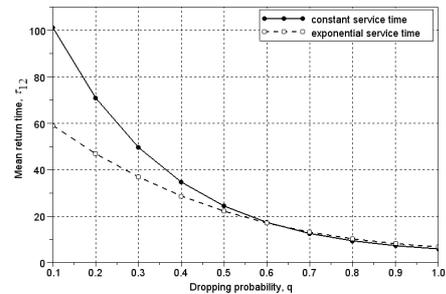


Figure 2: Dependence of mean return time τ_{12} from dropping probability q

header, and the packet is sent to the transport layer. The next layer Kernel Network Stack provides Packet Sending procedures. Once a packet completes processing at the SIP layer, it is passed on to the kernel for forwarding to the UAC/UAS.

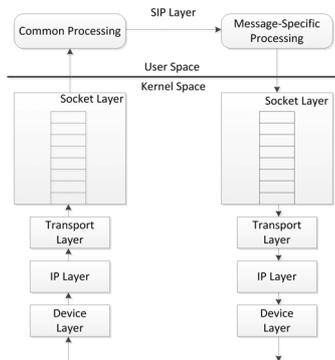


Figure 3: Message flow path through OS and SIP layers

First, for parsing, the text based SIP messages are syntactically analysed, broken down into parts, and converted into internal representations. Second, the newly created internal message representations have to be analyzed to infer on their later processing. For a SIP proxy server this means that it has to determine the messages destination and forward them towards there. For a UAS it means extracting the content of the message, probably displaying something to the user and creating and sending responses. Each of the two parts of processing a SIP message in a proxy server uses a specific amount of processing time. We assumed that parsing of a message always takes approximately the same amount of processing time, which independent of its type and content. The simulation model distinguishes requests and responses and assumes further that a response needs less processing time than a request. These assumptions lead to the proxy simulation model depicted in Fig. 4. Within this model, each message type refers to one of the four specific queues which are served by the processor P, according to priority scheduling (that means that a message from the FCFS queue with the highest priority is always taken first and forwarded to the processor; if there are no messages in the first queue left, a message with the next lower priority is taken, and so on). The following priorities are given to the message types (in decreasing order): Incoming (unparsed) external messages, Self-created re-transmissions, Parsed responses, Parsed requests.

A queueing Model (upper part) and a transaction manager (lower part) were added to the description above as main components. Upon the processing of a message, the SIP-specific transaction manager creates or deals with the corresponding state machine and returns the message into the queueing model for a second time, either as a re-transmission, request or response message.

Note that we assume that message reception and parsing is of highest priority, as it leads to the creation of the basic internal message representations in the mem-

ory. This model is based on an interrupt-driven system and therefore the network interface card triggers an interrupt upon reception of a message. That means further, the normal operation (that is SIP routing in terms of the simulation model) is interrupted as often as a new SIP message is received. After creating these structures, each message is again queued for routing until it is processed (routed). Within the routing process, the simulator creates transactions for each new request and they set timers which possibly create re-transmissions of requests which will then be sorted into the re-transmission queue.

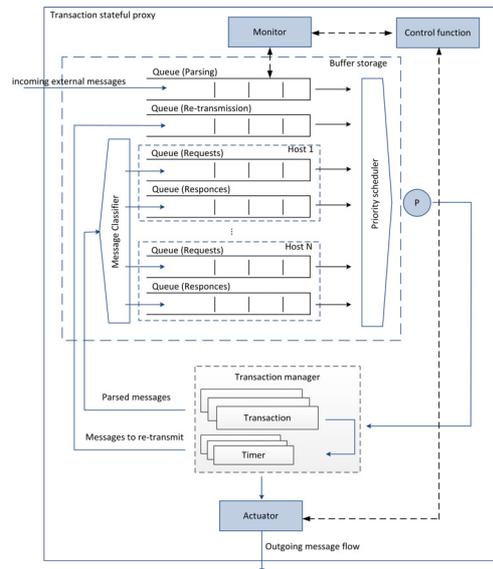


Figure 4: Message scheduling in proxy

CONCLUSION

In this paper we have considered two approaches for estimation performance measures of SIP server which operates in accordance with hysteretic overload control algorithm. We have built and analysed a queueing model with threshold control using the embedded Markov chain apparatus. The obtained formulas for calculation of the mean return time to the set of normal states are focused on numerical implementation. We have also presented the SIP simulator and considered several KPIs for its performance estimation.

Future work will be focused on mathematical modeling of overload control algorithms to make a proposal for the preliminary values of the control parameters, which will be used as input data for the simulator, and on investigation the impact of control parameters on the effectiveness of server performance. Development of the simulator will be continued to support the possibility of KPI's calculation and expansion of implemented FSM that includes different SIP call flow scenarios.

Notes and Comments. This work was supported in part by the Russian Foundation for Basic Research (grants 12-07-00108 and 13-07-00665).

REFERENCES

- Abaev, P., Gaidamaka, Yu., Samouylov, K. 2012. Modeling of Hysteretic Signaling Load Control in Next Generation Networks. Lecture Notes in Computer Science. Germany, Heidelberg, Springer-Verlag. –Vol. 7469. –P.371–378.
- Abaev, P., Gaidamaka, Yu., Samouylov, K. 2012. Queuing Model for Loss-Based Overload Control in a SIP Server Using a Hysteretic Technique. Lecture Notes in Computer Science. Germany, Heidelberg, Springer-Verlag. –Vol. 7469. –P.440–452.
- Abaev, P., Gaidamaka, Yu., Pechinkin, V., Razumchik, R., Shorgin, S. 2012. Simulation of overload control in SIP server networks. Proceedings of the 26th European Conference on Modelling and Simulation, ECMS 2012. –Germany, Koblenz. –Pp. 533–539.
- Abaev, P., Gaidamaka, Yu., Samouylov, K., Shorgin, S. 2013. Design And Software Architecture Of Sip Server For Overload Control Simulation. Proceedings of the 27th European Conference on Modelling and Simulation, ECMS 2013. –Norway, Alesund. –Pp. 580–586.
- Abaev, P., Pechinkin, V., Razumchik, R. 2012. On analytical model for optimal SIP server hop-by-hop overload control. Communications in Computer and Information Science: Modern Probabilistic Methods for Analysis of Telecommunication Networks. Germany, Heidelberg, Springer-Verlag. –Vol. 356. –P.1–10.
- Rosenberg, J., Schulzrinne, H., Camarillo, G. et al. 2002. SIP: Session Initiation Protocol. RFC 3261.
- Gurbani, V., Hilt, V., Schulzrinne, H. 2013. Session Initiation Protocol (SIP) Overload Control. draft-ietf-soc-overload-control-14.
- Hilt, V., Noel, E., Shen, C., Abdelal, A. 2011. Design Considerations for Session Initiation Protocol (SIP) Overload Control. RFC 6357.
- Noel, E., Williams, P. 2012. Session Initiation Protocol (SIP) Rate Control. draft-ietf-soc-overload-rate-control-03.
- Rosenberg, J. 2008. Requirements for Management of Overload in the Session Initiation Protocol. RFC 5390.
- Rosenberg, J. 2008. Session Initiation Protocol (SIP) Basic Call Flow Examples. RFC 3665.
- Gaidamaka, Yu., Pechinkin, A., Razumchik, R., Samouylov, K., Sopin, E. 2014. Analysis of M/G/1/R Queue with Batch Arrivals and Two Hysteretic Overload Control Policies. International Journal of Applied Mathematics and Computer Science (in print).
- Paxson, V., Almes, G., Mahdavi, J., Mathis, M. 1998. Framework for IP Performance Metrics. RFC 2330.
- Davis, M., Phillips, A., Umaoka, Y. 2010. BCP 47 Extension U. RFC 6067.
- Happenhofer, M., Egger, C., Reichl, P. 2010. Quality of Signaling: A new Concept for Evaluating the Performance of Non-INVITE SIP Transactions. Proc. of 22nd International Teletraffic Congress (ITC), 2010.
- Happenhofer, M., Reichl, P. 2010. Quality of Signaling (QoS) Metrics for Evaluating SIP Transaction Performance. In The 18th International Conference on Software, Telecommunications and Computer Networks. —Pp. 270-274.

AUTHOR BIOGRAPHIES

KONSTANTIN E. SAMOUYLOV received his Ph.D. from the Moscow State University and a Doctor of Sciences degree from the Moscow Technical University of Communications and Informatics. During 1985–1996 he held several positions at the Faculty of Sciences of the Peoples’ Friendship University of Russia where he became a head of the Telecommunication Systems Department in 1996. His current research interests are performance analysis of 3G networks, teletraffic of triple play networks, and signaling networks planning. He is the author of more than 100 scientific and technical papers and three books. His email address is ksam@sci.pfu.edu.ru.

PAVEL O. ABAEV received his Ph.D. in Computer Science from the Peoples’ Friendship University of Russia in 2011. He is an Associate Professor in the Telecommunication Systems department at Peoples Friendship University of Russia since 2013. His current research focus is on NGN signalling, QoS analysis of SIP, and mathematical modeling of communication networks. His email address is pabaev@sci.pfu.edu.ru.

YULIYA V. GAIDAMAKA received the Ph.D. in Mathematics from the Peoples’ Friendship University of Russia in 2001. Since then, she has been an associate professor in the university’s Telecommunication Systems department. She is the author of more than 50 scientific and conference papers. Her research interests include SIP signalling, multiservice and P2P networks performance analysis, and OFDMA based networks. Her email address is ygaidamaka@sci.pfu.edu.ru.

ALEXANDER V. PECHINKIN is a Doctor of Sciences in Physics and Mathematics and principal scientist at the Institute of Informatics Problems of the Russian Academy of Sciences, and a professor at the Peoples’ Friendship University of Russia. He is the author of more than 150 papers in the field of applied probability theory. His email address is apechinkin@ipiran.ru.

ROSTISLAV V. RAZUMCHIK received his Ph.D. in Physics and Mathematics in 2011. Since then, he has worked as a senior researcher at the Institute of Informatics Problems of the Russian Academy of Sciences. His current research activities focus on stochastic processes and queuing theory. His email address is rrazumchik@ieee.org

Simulation and Optimization

Regulation of the input flow of supply chains to optimize the production

C. D'Apice, C. De Nicola, R. Manzo

Department of Information Engineering, Electrical Engineering and Applied Mathematics
University of Salerno

84084, Fisciano, Salerno, Italy

Email: cdapice@unisa.it cdenicola@unisa.it rmanzo@unisa.it

KEYWORDS

Supply chains; Fluid dynamic model; Simulation; Optimization of input flow

ABSTRACT

The optimal control problem of adjusting the inflow, of piecewise constant type, to a supply chain in order to minimize the queue size and the quadratic difference between the outflow and the expected one is considered. The controls are represented by the duration of injections of different amounts of goods. The supply chain is modelled by a PDE-ODE: the conservation law describes the density of processed parts and the ODE the queue buffer occupancy. The numerical technique is based on the extensive use of generalized tangent vectors to a piecewise constant control, which represent time shifts of discontinuity points.

Introduction

The development of techniques for simulation and optimization purposes of industrial production is of great interest in order to answer questions raising in supply chain planning (optimal processing parameters, minimizing inventories to reduce costs or to ensure fully loaded production lines, and so on). Basically, we distinguish between steady state and instationary models which are time-dependent. A well-known class of stationary models are queuing theory models, which allow the calculation of several performance measures including the mean waiting time of goods in the system, the proportion of time the processors are busy and so on. In contrast, instationary models predict the time evolution of parts and include a dynamic inside the different production steps. The latter can be again divided into two classes: discrete (Discrete Event Simulations, [Forrester 1964]) or continuous (Differential Equations, [Armbruster et al. 2006], [Armbruster et al. 2007], [Helbing et al. 2004]). The latter class, thought in particular for large volume production on networks where a discrete description might fail, includes models based on partial differential equations ([Bretti et al. 2007], [D'Apice et al. 2010], [D'Apice and Manzo 2006], [D'Apice et al. 2009], [Göttlich et al. 2005]). Stochastic inputs to fluid dynamic models can be used to catch real behavior of complex systems.

In this paper, we focus on how to control in some case studies the flow through a supply chain so that a

desired amount of goods can be produced and storage costs are minimized. The starting point is a continuous model for supply chains proposed by Göttlich, Herty and Klar in [Göttlich et al. 2005], briefly GHK model. A supply chain consists of processors with constant processing rate and a queue in front of each processor. The dynamics of parts on a processor is described by a conservation law, while the evolution of the queue buffer occupancy is given by an ordinary differential equation, determined by the difference of fluxes between the preceding and following processors.

Various optimal control problems, corresponding to different types of controls, have been analysed for the GHK model (see [D'Apice et al. 2010], [D'Apice et al. 2011], [Göttlich et al. 2010], [Göttlich et al. 2010], [Herty and Klar 2003], [Kirchner et al. 2006]), such as the problem of determining optimal velocities for each individual processing unit or, in the case of networks with a vertex of dispersing type (splitting in more lines), the distribution rate has been controlled to minimize queues. In [D'Apice et al. 2011], piecewise constant controls are considered together with generalized tangent vectors, which represent time shifts of discontinuities of the control. The technique of such generalized tangent vectors was extensively used for conservation laws, see [Bressan et al. 2000], and for the case of network models, see [Garavello and Piccoli 2006], [Herty et al. 2007]. The main result of [D'Apice et al. 2011] is the existence of optimal controls. In [D'Apice et al. 2013], an innovative numerical approach, which builds up on the idea of generalized tangent vectors, is presented, in order to solve the optimal control problem, where the control is given by the input flow to the supply chain and the cost functional J is the sum of time-integral of queues and quadratic distance from a preassigned desired outflow. In particular the controls are the locations of the discontinuities of the input flow of piecewise constant type, while the flux values are fixed. The numerical method is based on perturbations of piecewise constant controls, obtained by time shifting the discontinuity points. Generalized tangent vectors have a particularly simple evolution in time, which can be conveniently adapted to easily implementable methods such as Upwind-Euler (briefly UE, see [Cutolo et al. 2011]). The discretization of the evolution of generalized tangent vectors allows the numerical computation of the cost functional gradient. This can be combined

with classical steepest descent or more advanced Newton methods for the optimization procedure by iterations.

The outline of the paper is the following. In the first section the supply chain model is described together with the optimal control problem. Then the Wave Front Tracking algorithm to construct approximate solutions to the model and the definition of generalized tangent vectors is briefly reported. A section is devoted to the UE numerical algorithm able to find solutions to the ODE-PDE system and also to the numerics for generalized tangent vectors and cost functional derivative. Finally the Euler-Upwind steepest descent algorithm is applied to a case study.

An optimal control problem for supply chains

A supply chain consists of suppliers. Each supplier is composed of a processor for parts assembling and construction. Each processor has an upper limit of parts that can be handled simultaneously. To avoid congestions, each processor has its own buffering area (queue), located in front of the processor, where the parts are possibly stored before the actual processing starts.

Formally a supply chain is a finite directed graph $G = (V, J)$ with arcs representing processors I_j , $j \in J = \{1, \dots, P\}$ and vertices, in $V = \{1, \dots, P - 1\}$, representing queues, in front of each processor, except the first. Each processor is parametrized by a bounded closed interval $I_j = [a_j, b_j]$, with $b_{j-1} = a_j$, $j = 2, \dots, P$. The maximal processing rate μ_j , and the processing velocity, $v_j = L_j/T_j$ with T_j and $L_j = b_j - a_j$ the processing time and the length of the j -th processor, are user-defined parameters on each arc. The evolution of density along the j -th processor is given by a conservation law

$$\partial_t \rho_j(x, t) + \partial_x f_j(\rho_j(x, t)) = 0, \quad \forall x \in [a_j, b_j], \quad \forall t > 0, \quad (1)$$

$$\rho_j(x, 0) = \rho_{j,0}(x), \quad \rho_j(a_j, t) = \frac{f_{j,inc}(t)}{v_j},$$

where the flux function $f_j(\rho_j(x, t))$ is given by:

$$f_j : [0, +\infty[\rightarrow [0, \mu_j], \\ f_j(\rho_j(x, t)) = \min\{\mu_j, v_j \rho_j(x, t)\}, \quad (2)$$

with $\rho_j \in [0, \rho_j^{max}]$ the unknown function, representing the density of parts, while the initial datum $\rho_{j,0}$ and the inflow $f_{j,inc}(t)$ have to be assigned. An input profile $u(t)$ on the left boundary $\{(a_1, t) : t \in \mathbb{R}\}$ is given for the first arc of the supply chain. Each queue buffer occupancy is modelled as a time-dependent function $t \rightarrow q_j(t)$. The dynamics of the buffering queue is governed by the following equation:

$$\dot{q}_j(t) = f_{j-1}(\rho_{j-1}(b_{j-1}, t)) - f_{j,inc}, \quad j = 2, \dots, P, \quad (3)$$

where the first term is defined by the trace of ρ_{j-1} (which is assumed to be of bounded variation on the x variable), while the second is defined by:

$$f_{j,inc} = \begin{cases} \min\{f_{j-1}(\rho_{j-1}(b_{j-1}, t)), \mu_j\} & \text{if } q_j(t) = 0, \\ \mu_j & \text{if } q_j(t) > 0. \end{cases} \quad (4)$$

This allows the following interpretation. We process as many parts as possible. If the outgoing buffer is empty, then we process all incoming parts but at most μ_j , otherwise we can always process at rate μ_j . Summarizing, we obtain a system of partial differential equations governing the dynamics on each processor coupled to ordinary differential equations for the evolution of queues:

$$\begin{cases} \partial_t \rho_j(x, t) + \partial_x \min\{\mu_j, v_j \rho_j(x, t)\} = 0 & j = 1, \dots, P, \\ \dot{q}_j(t) = f_{j-1}(\rho_{j-1}(b_{j-1}, t)) - f_{j,inc}(t) & j = 2, \dots, P, \\ q_j(0) = q_{j,0} & j = 2, \dots, P, \\ \rho_j(x, 0) = \rho_{j,0}(x) & j = 1, \dots, P, \\ \rho_j(a_j, t) = \frac{f_{j,inc}(t)}{v_j} & j = 1, \dots, P, \\ f_{1,inc}(t) = u(t) & \end{cases} \quad (5)$$

where $f_{j,inc}$ is given by (4), for $j = 2, \dots, P$.

When we fix a time horizon $[0, T]$, we can define the cost functional:

$$\begin{aligned} J(u) &= \sum_{j=2}^P \int_0^T \alpha_1(t) q_j(t) dt + \\ &+ \int_0^T \alpha_2(t) [v_P \cdot \rho_P(b_P, t) - \psi(t)]^2 dt \\ &\doteq J_1(u) + J_2(u), \end{aligned} \quad (6)$$

where α_1, α_2 are weight functions, (ρ_j, q_j) is the solution to (5) for the control u , $v_P \cdot \rho_P(b_P, t)$ represents the outflow of the supply chain (assuming the density level is below μ_P), while $\psi(t) \in \mathbb{R}$ is a pre-assigned desired outflow. Given $C > 0$, we consider the minimization problem

$$\min_{u \in \mathcal{U}_C} J(u) \quad (7)$$

where $\mathcal{U}_C = \{u : [0, T] \rightarrow [0, \mu_1]; u \text{ measurable, } T.V.(u) \leq C\}$ (with $T.V.$ indicating the total variation). In other words, we want to minimize the queues length and the distance between the exiting flow and the pre-assigned flow $\psi(t)$, using the supply chain input u as control. In particular, we assume an input flow of piecewise constant type, and fixing the levels we control the discontinuities times, it means that we aim to regulate the injection times of the different amount of goods.

In [D'Apice et al. 2011], the existence of an optimal control was proved for a general problem, which includes the case (5)-(7), while in [D'Apice et al. 2013] a new approach to solve (5)-(7) numerically is provided. The key idea is to focus on piecewise constant controls and perturb the position of discontinuity points. The procedure corresponds to define (generalized) tangent vectors to u (in the spirit of [Bressan et al. 2000]), taking advantage of the knowledge of time evolution of such tangent vectors, developed in [Herty et al. 2007]. For every $u \in \tilde{\mathcal{U}} \subset \mathcal{U}_C$, with \mathcal{U}_C the set of Piecewise Constant controls we indicate by $\tau_k = \tau_k(u)$, $k = 1, \dots, \delta(u)$, the discontinuity points of u (see 1). The perturbation to a piecewise constant control is defined as follows. Given $u \in \tilde{\mathcal{U}}$, a tangent vector to u is a vector $\xi = (\xi_1, \dots, \xi_{\delta(u)}) \in \mathbb{R}^{\delta(u)}$ representing shifts of discontinuities. The norm of the tangent vector is

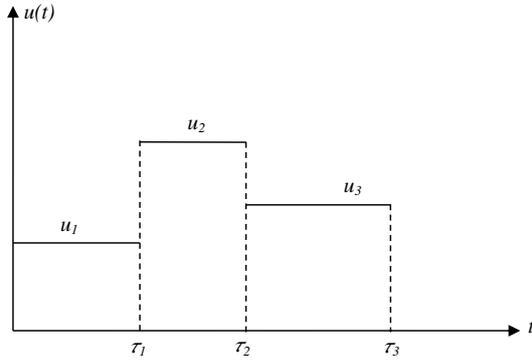


Fig. 1. Input flow.

defined as:

$$\|\xi\| = \sum_{k=1}^{\delta(u)} |\xi_k| \cdot |u(\tau_k+) - u(\tau_k-)|,$$

where $(\cdot +)$ and $(\cdot -)$ indicate the limit from the right and the left. Assume now for simplicity that $\tau_1 > 0$, $\tau_{\delta(u)} < T$, and set $\tau_0 = 0$, $\xi_0 = 0$, $\tau_{\delta(u)+1} = T$, $\xi_{\delta(u)+1} = 0$. Then given a tangent vector ξ to u , for every ε sufficiently small we define the infinitesimal displacement as:

$$u_\varepsilon = \sum_{k=0}^{\delta(u)} \chi_{[\tau_k + \varepsilon \xi_k, \tau_{k+1} + \varepsilon \xi_{k+1}]} [u(\tau_k+)], \quad (8)$$

where χ is the indicator function. In other words u_ε is obtained from u by shifting the discontinuity points of $\varepsilon \xi$.

In the next sections numerical schemes for the solution of (5) and for the evolution of tangent vectors will be defined. The latter, in turn, will provide the information for the computation of numerical gradient of the cost functional J .

The evolution of tangent vectors is particularly clear for the theoretical numerical scheme given by the WFT algorithm. We thus recall briefly the WFT algorithm and the evolution of tangent vectors along approximate solutions constructed via the WFT algorithm.

The Wave Front Tracking algorithm

In this section we explain how to construct piecewise constant approximate solutions to (5) by WFT method, see [Bressan 2000] for details.

Given a discretization parameter σ and initial conditions $\rho_{j,0}$ in BV, the space of bounded variation functions, a WFT approximate solution is constructed by a procedure sketched by the following steps:

- Approximate the initial datum by a piecewise constant function (with discretization parameter σ) and solve the Riemann Problems (RPs) corresponding to discontinuities of the approximation. In RPs solutions approximate rarefactions by rarefaction shocks of size σ ;
- Use the piecewise constant solution obtained piecing together the solutions to RPs up to the first time of interaction of two shocks;

- Then solve the new RP created by interaction of waves and prolong the solution up to next interaction time, and so on.

Notice that, as soon as a boundary datum will achieve a value below μ_j , then in finite time all values above μ_j will disappear from the j -th processor, see also [Herty et al. 2007]. Therefore, for simplicity, we will assume (H1) $\rho_{j,0}(x) \leq \mu_j$ for all $j \in \mathcal{J}$.

Then the same inequality will be satisfied for all times. In this case solutions to RPs are particularly simple, indeed the conservation law is linear, thus given some Riemann data (ρ_-, ρ_+) on the j -th processor, the solution is always given by a shock travelling with velocity v_j .

Tangent vectors evolution

The infinitesimal displacement of each discontinuity of the control u produces changes in the whole supply chain, whose effects are visible both on processors and on queues. In fact, every shift ξ generates shifts on the densities and shifts on the queues, which spread along the whole supply chain.

A tangent vector to the solution (ρ_j, q_j) to (5) is given by:

$$({}^\beta \xi_j, \eta_j),$$

where β runs over the set of discontinuities of ρ_j , ${}^\beta \xi_j$ are the shifts of the discontinuities, while η_j is the shift of the queue buffer occupancy q_j . The norm of a tangent vector is given by:

$$\|({}^\beta \xi_j, \eta_j)\| = \sum_{\beta} |{}^\beta \xi_j| |\Delta^\beta \rho_j| + \sum_j |\eta_j|, \quad (9)$$

where $\Delta^\beta \rho_j = {}^\beta \rho_j^l - {}^\beta \rho_j^r$ is the jump in ρ of the discontinuity β (where ${}^\beta \rho_j^l$, respectively ${}^\beta \rho_j^r$, is the value on the left, respectively right, of the discontinuity β). Because of assumption (H1), we have no wave interaction inside the processors. Therefore, densities and queues shifts remain constant for almost all times and change only at those times at which one of the following interactions occurs:

- interaction of a density wave with a queue;
- emptying of the queue.

Assume a wave with shift ${}^\beta \xi_{j-1}$ interact with the j -th queue and let \bar{t} be the interaction time. The symbols $+$ and $-$ indicate quantities before and after \bar{t} , respectively. So, ρ_j^- and ρ_j^+ indicate the densities on the processor I_j before and after an interaction occurs and similarly for I_{j-1} . Moreover ${}^\beta \xi_j$ denote the shift on the processor I_j and with ${}^\beta \eta_j^-$ and ${}^\beta \eta_j^+$ the shifts on the queue q_j , respectively before and after the interaction. In case **a)** two subcases have to be distinguished:

- $q_j(\bar{t}) = 0$;
- $q_j(\bar{t}) > 0$.

In case **a.1)** two further subcases have to be considered:

- a.1.1)** if $v_{j-1} \rho_{j-1}^+ < v_{j-1} \rho_{j-1}^- < \mu_j$, then ${}^\beta \xi_j = \frac{v_j}{v_{j-1}} {}^\beta \xi_{j-1}$ and ${}^\beta \eta_j^- = 0 = {}^\beta \eta_j^+$;
- a.1.2)** if $v_{j-1} \rho_{j-1}^+ > \mu_j$, then ${}^\beta \xi_j = \frac{v_j}{v_{j-1}} {}^\beta \xi_{j-1}$ and ${}^\beta \eta_j^+ = {}^\beta \xi_{j-1} \frac{(v_{j-1} \rho_{j-1}^+ - \mu_j)}{v_{j-1}} + {}^\beta \eta_j^-$.

In case **a.2**) we have: ${}^\beta\xi_j = 0$, ${}^\beta\eta_j^+ = {}^\beta\xi_{j-1}(\rho_{j-1}^- - \rho_{j-1}^+) + {}^\beta\eta_j^-$.

Finally in case **b**) we get: ${}^\beta\eta_j^+ = 0$, ${}^\beta\xi_{j-1} = 0$ and ${}^\beta\xi_j = -\frac{v_j {}^\beta\eta_j^-}{(v_{j-1}\rho_{j-1}^- - \mu_j)}$. Using the above notations, we indicate with ${}^\beta\xi_P$ the shift to a generic discontinuity of ρ_P and with ${}^\beta\rho_P^+$, respectively ${}^\beta\rho_P^-$, the value of ρ_P on the right, respectively left, of the discontinuity. Considering a control $u \in \tilde{\mathcal{U}}$ and a tangent vector $\xi \in \mathbb{R}^{\delta(\cong)}$ to u , the gradient of the cost functional J with respect to ξ is given by:

$$\nabla_\xi J(u) = Y_1^{WFT} + Y_2^{WFT}, \quad (10)$$

where

$$Y_1^{WFT} \doteq \sum_j \int_0^T \alpha_1(t) \eta_j(t) dt,$$

$$Y_2^{WFT} \doteq \sum_\beta [\alpha_2(t^\beta) v_P ({}^\beta\rho_P^+ + {}^\beta\rho_P^- - 2\psi(t^\beta)) {}^\beta \cdot \xi_P \Delta({}^\beta\rho_P)]$$

with $\Delta({}^\beta\rho_P) = {}^\beta\rho_P^+ - {}^\beta\rho_P^-$ and t^β the interaction time of the discontinuity indexed by β with b_P , the right extreme of the supply chain.

Steepest descent for the Upwind-Euler scheme

In this section an Upwind-Euler scheme for the system (5) and then a numerical scheme for the evolution of the tangent vectors to a solution to the PDE-ODE model are introduced. From the latter it is possible to compute numerically the derivative of the cost functional with respect to the discontinuities of the input flow. This, in turn, will be used in steepest descent methods to find the optimal control.

For simplicity we assume:

(H2) The lengths L_j are rationally dependent.

Assumption (H2) allows us to use a unique space mesh for all processors I_j , $j = 1, \dots, P$. Indeed there exists Δ so that all L_j are multiple of a value Δ and we will always use time and space meshes dividing by Δ .

In the next section briefly the Upwind-Euler method, analysed in [Cutolo et al. 2011] to construct numerical solutions to the supply chain model (5) is reported.

A. Upwind-Euler scheme for supply chains

Given a space mesh Δx , for each processor I_j , we set $\Delta t_j = \Delta x/v_j$ and define a numerical grid of $[0, L_j] \times [0, T]$ by:

- $(x_i, t^n)_j = (i\Delta x, n\Delta t_j)$, $i = 0, \dots, N_j$, $n = 0, \dots, M_j$ are the grid points;
- ${}^j\rho_i^n$ is the value taken by the approximated density at the point $(x_i, t^n)_j$;
- q_j^n is the value taken by the approximate queue buffer occupancy at time t^n .

The Upwind method reads:

$${}^j\rho_i^{n+1} = {}^j\rho_i^n - \frac{\Delta t_j}{\Delta x} v_j ({}^j\rho_i^n - {}^j\rho_{i-1}^n) = {}^j\rho_{i-1}^n, \quad (11)$$

where $j \in \mathcal{J}$, $i = 0, \dots, N_j$ and $n = 0, \dots, M_j$. Notice that the CFL condition is given by $\Delta t_j \leq \frac{\Delta x}{v_j}$, and thus holds true. The explicit Euler method is given by:

$$\begin{aligned} q_j^{n+1} &= q_j^n + \Delta t_j (f_{j-1}^n - f_{j,inc}^n), \quad j \in \mathcal{J} \setminus \{\infty\}, \\ n &= 0, \dots, M_j, \end{aligned} \quad (12)$$

where f_{j-1}^n needs to be defined and

$$f_{j,inc}^n = \begin{cases} \min\{f_{j-1}^{(j-1)\rho_{N_{j-1}}^n}, \mu_j\} & q_j^n(t) = 0, \\ \mu_j & q_j^n(t) > 0. \end{cases} \quad (13)$$

Now, if $\Delta t_{j-1} \leq \Delta t_j$ we set:

$$\begin{aligned} f_{j-1}^n &= \sum_{l=1}^{M(n)-m(n)-1} \Delta t_{j-1} f_{j-1}^{(j-1)\rho_{N_{j-1}}^{m(n)+l}} = \\ &= \sum_{l=1}^{\gamma} \Delta t_{j-1} f_{j-1}^{(j-1)\rho_{N_{j-1}}^{\gamma n+l}}, \end{aligned} \quad (14)$$

where $m(n)$ and $M(n)$ are defined as:

$$m(n) = \sup\{m : m\Delta t_{j-1} \leq n\Delta t_j\},$$

$$M(n) = \inf\{M : M\Delta t_{j-1} \geq (n+1)\Delta t_j\}.$$

Otherwise, that is if $\Delta t_{j-1} > \Delta t_j$, we set:

$$f_{j-1}^n = f_{j-1} \left(j-1 \rho_{N_{j-1}}^{\lfloor \frac{n\Delta t_j}{\Delta t_{j-1}} \rfloor} \right), \quad (15)$$

where $\lfloor \cdot \rfloor$ indicates the floor function. Boundary data are treated using ghost cells and the expression of in-flows given by (13). The convergence of the scheme has been proved in [Cutolo et al. 2011] using a comparison with WFT approximate solutions.

B. Numerics for tangent vectors and cost functional

First the control space is discretized via the time mesh Δt :

$$\begin{aligned} \tilde{\mathcal{U}}_{\Delta t} &= \{u \in \tilde{\mathcal{U}} : \tau_k(u) = n(u, k) \Delta t, \quad n(k, u) \in \mathcal{N}, \\ &\quad k = 1, \dots, \delta(u)\}. \end{aligned}$$

Now for every $u \in \tilde{\mathcal{U}}_{\Delta t}$ shifts ξ are considered so that the obtained time-shifted control is still in $\tilde{\mathcal{U}}_{\Delta t}$. Then every ξ_k is necessarily a multiple of Δt . Hence from now on we will restrict to the case:

$$\xi_k = \pm \Delta t, \quad k = 1, \dots, \delta(u).$$

For a generic processor I_j and a discontinuity point τ_k of the control, ${}^{k,j}\xi_i^n$ and ${}^{k,j}\eta^n$ denote the approximations of ${}^k\xi_j(x_i, t^n)$, and ${}^k\eta_j(t^n)$, respectively. Such approximations are defined by a recursive procedure explained in the following.

The tangent vector approximations are initialized by setting:

$$\begin{aligned} {}^{k,j}\xi_i^n &= 0, \quad \text{for } n = 1, \dots, n(k-1, u), \\ j &= 1, \dots, P, \\ {}^{k,1}\xi_1^{n(k,u)} &= v_1(\pm \Delta t), \\ {}^{k,j}\eta^0 &= 0, \quad j = 1, \dots, P. \end{aligned}$$

The definition of $k,1\xi_1^{n(k,u)}$ reflects the fact that the shift ξ_k provokes a shift of the wave generated on the first processor.

Now, the evolution of approximations of tangent vectors to ρ inside processors is simply given by:

$$k,j\xi_i^{n+1} = k,j\xi_{i-1}^n.$$

On the other side, the approximation of ξ and η influence each other at interaction times with queues. More precisely, the four cases described in the previous section are considered, obtaining:

$$a.1.1): k,j\eta^{n+1} = 0, k,j\xi_1^{n+1} = \frac{v_j}{v_{j-1}} k,j-1\xi_{N_{j-1}}^n;$$

$$a.1.2): k,j\xi_1^{n+1} = \frac{v_j}{v_{j-1}} k,j-1\xi_{N_{j-1}}^n, k,j\eta^{n+1} =$$

$$= k,j-1\xi_{N_{j-1}}^n \frac{(v_{j-1}^{j-1}\rho_{N_{j-1}}^{n+1} - \mu_j)}{v_{j-1}} + k,j\eta^n;$$

$$a.2): k,j\xi_1^{n+1} = 0, k,j\eta^{n+1} =$$

$$= k,j-1\xi_{N_{j-1}}^n \left(j-1\rho_{N_{j-1}}^{n+1} - j-1\rho_{N_{j-1}}^n \right) + k,j\eta^n;$$

$$b): k,j-1\xi_{N_{j-1}}^n = 0, k,j\eta^{n+1} = 0, k,j\xi_1^{n+1} =$$

$$- \frac{v_j k,j\eta^n}{v_{j-1} j-1\rho_{N_{j-1}}^n - \mu_j}.$$

Now numerical approximations for $\nabla_\xi J$ can be computed. Denote by k,jY_1^n , respectively kY_2^n , the numerical approximations of the k -th component of Y_1^{WFT} , respectively Y_2^{WFT} , as defined in (10) on processor I_j at time t^n .

Such approximation is initialized by setting:

$$k,jY_1^0 = 0, kY_2^0 = 0, \quad j = 1, \dots, P, \quad k = 1, \dots, \delta(u).$$

The evolution is determined by the following simple rules. For Y_1 , if $q_j^{n+1} > 0$, then we set

$$k,jY_1^{n+1} = k,jY_1^n + \alpha_1(t^n) k,j\eta^n \Delta t,$$

while for $q_j^{n+1} = 0$ two subcases are distinguished:

- if $q_j^n = 0$, then $k,jY_1^{n+1} = k,jY_1^n$;
- if $q_j^n > 0$, then $k,jY_1^{n+1} = k,jY_1^n + \frac{1}{2} \alpha_1(t^n) k,j\xi_1^{n+1} k,j\eta^n$.

For Y_2 we set:

$$kY_2^{n+1} = kY_2^n + \alpha_2(t^n) v_P \left(\left(\rho_{N_P}^n - \psi(t^n) \right)^2 - \left(\rho_{N_P}^{n+1} - \psi(t^n) \right)^2 \right) k,P\xi_{N_P}^n.$$

A steepest descent algorithm, denoting with ϑ the iteration step, is defined by setting

$$\tau_k^{\vartheta+1} = \tau_k^{\vartheta} + \left\lfloor \frac{h_\theta \left(\sum_j \sum_n k,jY_1^n + \sum_n kY_2^n \right)}{\Delta t} \right\rfloor \Delta t,$$

where h_θ is a coefficient to be suitably chosen. More precisely the parameter h_θ may be chosen to solve an optimization problem to get specific schemes. In [D'Apice et al. 2013] convergence results and error estimates for the Upwind-Euler steepest descent scheme are provided.

Simulations

In this Section we use the Euler-Upwind steepest descent algorithm on a test case. The latter concerns a painting process of the cars bonnet, which consists of the following workflow: the bonnets enter the process and are painted, then they pass to the cooking stage after that they are washed. Finally, the bonnets are packaged and distributed to the retailers. This process can be represented by a sequential graph where the vertices summarize the activities and arcs the transitions between different activities, as shown in the following figure. Consider the following input function:

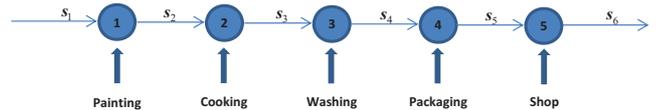


Fig. 2. Painting process of bonnett.

$$u(t) = \begin{cases} u_1 & 0 \leq t \leq \tau_1, \\ u_2 & \tau_1 < t \leq \tau_2, \\ u_3 & \tau_2 < t \leq T. \end{cases}$$

and fix the wished output flow $\psi(t) = 50$.

We assume that the supply chain (see Figure 2) is initially empty, $v_j = 1$ for every $j = 1, \dots, 6$ and

$$\mu_1 = 100, \mu_2 = 70, \mu_3 = 50, \mu_4 = 30, \mu_5 = 40, \mu_6 = 20.$$

For simplicity we set $\alpha_1 \equiv 1$ and $\alpha_2 \equiv 1$ and analyze two different cases:

Case a) $u_1 = 80, u_2 = 90, u_3 = 110, T = 35$, initial values $(\tau_1, \tau_2) = (2, 5)$;

Case b) $u_1 = 100, u_2 = 50, u_3 = 80, T = 35$, initial values $(\tau_1, \tau_2) = (2, 5)$.

As time and space grid meshes we choose $\Delta x = 0.02$ and $\Delta t = 0.016$ so as to satisfy the CFL condition. We use the condition that J remains unchanged for five runnings of the algorithm as forced stop criterion. The times τ_1 and τ_2 found by the algorithm are:

Case a) $\tau_1 \simeq 33.99, \tau_2 \simeq 34.15$: as expected both τ_1 and τ_2 run toward T ; in fact, in order to minimize the queues, the optimization algorithm tends to reduce the inflow levels which increase the queues (i. e. u_2 and u_3).

Case b) $\tau_1 = 0, \tau_2 \simeq 33.91$: τ_1 runs to zero and τ_2 runs toward T ; as in the previous case, the optimization algorithm works to reduce the inflow levels which lead to queues increasing (i. e. u_1 and u_3).

In Table I we report the numerical values of τ_1, τ_2 and J at each iteration of the steepest descent algorithm for Case a).

Figures 3-4 depict the values assumed by J , and (τ_1, τ_2) , at each iteration step for Case a). The behaviour of the cost functional J in the plane (τ_1, τ_2) is reported for Case b) in Figure 5, to confirm the goodness of the steepest descent algorithm.

Iteration	τ_1	τ_2	J
1	2	5	6259043.87
2	8.39	16.58	5975423.18
3	13.50	23.53	5840830.02
4	17.59	27.70	5771879.74
5	20.87	30.20	5733613.88
6	23.49	31.70	5711361.84
7	25.58	32.61	5697897.12
8	27.26	33.15	5689631.94
9	28.60	33.48	5684376.46
10	29.67	33.67	5681060.69
11	30.53	33.79	5678934.30
12	31.21	33.86	5677581.40
13	31.76	33.90	5676726.63
14	32.20	33.93	5676179.55
15	32.55	33.95	5675820.15
16	32.84	33.96	5675585.43
17	33.06	33.97	5675437.89
18	33.24	33.98	5675344.61
19	33.39	33.98	5675282.03
20	33.50	33.99	5675237.87
21	33.59	34.00	5675214.97
22	33.67	34.00	5675196.57
23	33.72	34.00	5675187.45
24	33.77	34.01	5675179.41
25	33.80	34.02	5675175.09
26	33.82	34.03	5675171.17
27	33.84	34.05	5675169.43
28	33.85	34.06	5675169.43
29	33.86	34.07	5675167.90
30	33.88	34.07	5675166.61
31	33.89	34.08	5675165.43
32	33.89	34.09	5675165.43
33	33.90	34.10	5675164.42
34	33.91	34.10	5675164.42
35	33.92	34.10	5675163.58
36	33.92	34.11	5675163.57
37	33.93	34.11	5675163.57
38	33.94	34.12	5675162.87
39	33.94	34.12	5675162.87
40	33.94	34.12	5675162.87
41	33.95	34.12	5675162.30
42	33.95	34.13	5675162.30
43	33.96	34.13	5675162.30
44	33.96	34.13	5675162.30
45	33.96	34.13	5675161.85
46	33.97	34.13	5675161.85
47	33.97	34.14	5675161.85
48	33.97	34.14	5675161.85
49	33.97	34.14	5675161.85
50	33.97	34.14	5675161.85
51	33.98	34.14	5675161.85
52	33.98	34.14	5675161.50
53	33.98	34.14	5675161.50
54	33.98	34.14	5675161.50
55	33.99	34.14	5675161.50
56	33.99	34.14	5675161.50
57	33.99	34.15	5675161.50

TABLE I: Case *a*: values of τ_1 , τ_2 , and corresponding J in 57 iterations of the steepest descent algorithm.

Notice that since J decreases when the number of iteration increases, the optimization of τ_1 and τ_2 allows an effective decrement of queues sizes.

The values of τ_1 and τ_2 , found by the steepest descent algorithm, seem to be grid independent; in fact choosing different time and space grid meshes the variation of the optimal discontinuities values is not meaningful: for $\Delta x = 0.01$ and $\Delta t = 0.008$ the algorithm stops with values of (τ_1, τ_2) in a small neighborhood of $(33.99, 34.16)$ in 109 iterations, respectively $(0, 33.93)$ in 77 iterations, for Case *a*, respectively Case *b*.

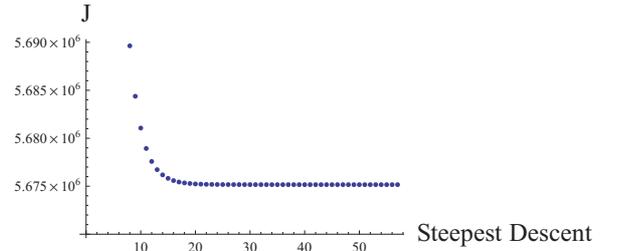


Fig. 3. Case *a*. J versus iteration steps.

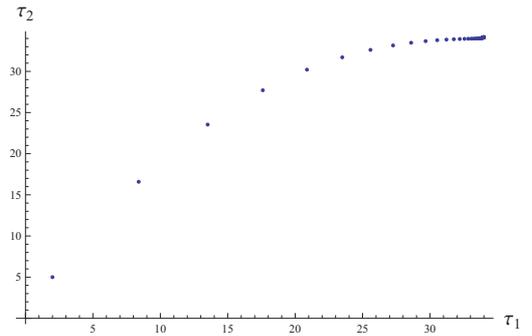


Fig. 4. Case *a*. “path” followed by the steepest descent algorithm in the plane (τ_1, τ_2) .

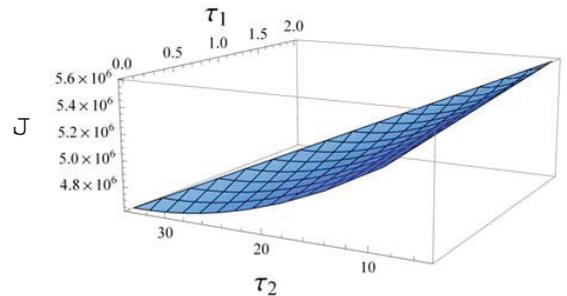


Fig. 5. Case *b*. Behaviour of J in the plane (τ_1, τ_2) .

CONCLUSIONS

In this paper, starting by a fluid dynamic model for supply chains, we used a new optimization method based on tangent vectors technique to find the best configuration of input flow in order to minimize the queues formation (representing bottlenecks for the sup-

ply chain) at each processor, and improve the final production. Due to the mathematical difficulties to solve analytically the PDE and ODE equations, we implemented a numerical method using the steepest descent algorithm to get our goal and applied it to some test cases. A further step forward, as future development, it could be done dealing with more complex supply network, where each node could be affected by multiple processors. In this case, incoming and outgoing flows should be considered, and some distribution rules should be implemented.

REFERENCES

- D. ARMBRUSTER, P. DEGOND AND C. RINGHOFER, *A model for the dynamics of large queueing networks and supply chains*, SIAM Journal on Applied Mathematics, 66(3) (2006), pp. 896–920.
- D. ARMBRUSTER, P. DEGOND AND C. RINGHOFER, *Kinetic and fluid models for supply chains supporting policy attributes*, Bulletin of the Inst. Math., Accademia Sinica, 2 (2007), pp. 433–460.
- A. BRESSAN, *Hyperbolic Systems of Conservation Laws - The One-dimensional Cauchy Problem*, 250 pages, Oxford Univ. Press, 2000.
- A. BRESSAN, G. CRASTA AND B. PICCOLI, *Well-Posedness of the Cauchy Problem for $n \times n$ Systems of Conservation Laws*, *Memoirs of the American Mathematical Society*, 146 (694), 2000.
- G. BRETTI, C. D’APICE, R. MANZO AND B. PICCOLI, *A continuum-discrete model for supply chains dynamics*, Netw. Heterog. Media, 2(4) (2007), pp. 661–694.
- A. CUTOLO, B. PICCOLI AND L. RARITÀ, *An Upwind-Euler scheme for an ODE-PDE model of supply chains*, Siam Journal on Scientific Computing, 33(4) (2011), pp. 1669–1688.
- C. D’APICE, S. GOETTICH, H. HERTY AND B. PICCOLI, *Modeling, Simulation and Optimization of Supply Chains*, 216 pages, SIAM, Philadelphia PA, USA, 2010.
- C. D’APICE AND R. MANZO, *A fluid dynamic model for supply chains*, Netw. Heterog. Media, 1(3) (2006), pp. 379–398.
- C. D’APICE, R. MANZO AND B. PICCOLI, *Modelling supply networks with partial differential equations*, Quarterly of Applied Mathematics, 67(3) (2009), pp. 419–440.
- C. D’APICE, R. MANZO AND B. PICCOLI, *Optimal input flows for a PDE-ODE model of supply chains*, Communication in Mathematical Sciences, 10(36) (2012), pp. 1226–1240.
- C. D’APICE, R. MANZO AND B. PICCOLI: *Numerical schemes for the optimal input flow of a supply-chain*, *SIAM Journal on Numerical Analysis*, 51(5) (2013), pp. 2634–2650.
- J.W. FORRESTER, *Industrial dynamics*, 464 pages. MIT Press, Cambridge, MA, 1964.
- M. GARAVELLO AND B. PICCOLI, *Traffic flows on networks*, American Institute of Mathematical Sciences, 243 pages, Springfield, MO, 2006.
- S. GÖTTLICH, M. HERTY AND A. KLAR, *Network models for supply chains*, *Communication in Mathematical Sciences*, 3(4) (2005), pp. 545–559.
- S. GÖTTLICH, M. HERTY AND A. KLAR, *Modelling and Optimization of Supply Chains on Complex Networks*, *Communication in Mathematical Sciences*, 4(2) (2006), pp. 315–330.
- S. GÖTTLICH, M. HERTY AND C. RINGHOFER, *Optimization of order policies in supply networks*, *European J. Oper. Res.*, 202(2) (2010), pp. 456–465.
- M. HERTY AND A. KLAR, *Modeling, Simulation, and Optimization of Traffic Flow Networks*, *SIAM J. Sci. Comput.*, 25(3) (2003), pp. 1066–1087.
- M. HERTY, A. KLAR AND B. PICCOLI, *Existence of solutions for supply chain models based on partial differential equations*, *SIAM J. Math. An.*, 39(1) (2007), pp. 160–173.
- D. HELBING, S. LÄMMER, T. SEIDEL, P. SEBA AND T. PLATKOWSKI, *Physics, stability and dynamics of supply networks*, *Physical Review E* 70, 066116 (2004).
- C. KIRCHNER, M. HERTY, S. GÖTTLICH AND A. KLAR, *Optimal control for continuous supply network models*, *Netw. Heterog. Media*, 1(4) (2006), pp. 675–688.

CIRO D’APICE is full professor in Mathematical Analysis at the Department of Information Engineering, Electrical Engineering and Applied Mathematics of the University of Salerno, Italy. He is the Director of two research centers: CRMPA - Centro di Ricerca in Matematica Pura ed Applicata, CEMSAC - Centro di Eccellenza su Metodi e Sistemi per Aziende Competitive. He is author of 2 books and more than 150 papers about homogenization and optimal control; conservation laws models for vehicular traffic, telecommunications and supply chains; spatial behaviour for dynamic problems; queueing systems and networks. His e-mail address is cdapice@unisa.it.

CARMINE DE NICOLA is an Electronics Engineer and received his Ph.D. in Mathematics in 2011, at University of Salerno. At the moment he is assistant researcher at University of Salerno, Italy. His current research activities focus on queuing theory, operation research and mathematical modeling of complex systems. His email address is cdenicola@unisa.it.

ROSANNA MANZO is a researcher in Mathematical Analysis at the Department of Information Engineering, Electrical Engineering and Applied Mathematics of the University of Salerno, Italy. She received PhD in Information Engineering from University of Salerno. Her research areas include fluid - dynamic models for traffic flows on road, telecommunication and supply networks, optimal control, queueing theory, self - similar processes, computer aided learning. She is author of about 50 papers appeared on international journals and many publications on proceedings. Her e-mail address is rmanzo@unisa.it.

SEASONAL TRENDS IN SUPPLY CHAINS

Hans-Peter Barbey
University of Applied Sciences Bielefeld
Wilhelm-Bertelsmann-Str. 10, 33602 Bielefeld, Germany
Email: hans-peter.barbey@fh-bielefeld.de

KEYWORDS

Supply chain, bullwhip effect, simulation, closed-loop control, order strategies.

ABSTRACT

Supply chains in industry have a very complex structure. The influence of many parameters is not known. Therefore the control of the material flow is rather difficult. In order to recognize these relationships between the parameters, initially very simple models have to be established. For this examination a linear supply chain with 4 stages will be designed. In all stages the stock is closed-loop controlled to a nominal stock. In particular, the seasonal trend on the control is considered. Therefore the only decision which can be done in the entire supply chain is the quantity of an order. It will be shown that the period of the trend has a significant influence to the quality of the controlling. A good quality of control results only for long periods of the seasonal trend.

1 INTRODUCTION

Dynamic behavior of the material flow in a supply chain is influenced by the order policy of each particular company of a supply chain. The interaction of all companies creates the bullwhip effect, which has been described first by (Forrester 1958). It is the increasing of a small variation in the requirements of a customer to an enormous oscillation with the manufacturer at the beginning of a supply chain. In many articles, this phenomenon is only described in general terms without a mathematical definition (i.e. Erlach 2010 and Dickmann 2007). It is questionable if the bullwhip effect can be avoided at all (Bretzke 2008). A mathematical justification for this thesis is not given in this paper. The main influences of the bullwhip effect are as follows (Gudehus 2005):

- Independent orders of the particular companies in a supply chain
- Synchronic orders (i.e. subsidiaries of one company)
- Wrong order policy in an emergency case

- Speculative order policy or sale actions

To avoid the bullwhip effect, cooperation between all members in a supply chain is necessary. Basically, informations about i.e. orders of customers have to be provided to all subsuppliers in the supply chain. This kind of cooperation is rather difficult in reality. The question is if the bullwhip effect can be avoided without any cooperation and providing of information to all members in a supply chain.

A very simple model of a supply chain has been published on the ECMS2013 (Barbey 2013). The target of this simulation was to develop strategies for a closed-loop control of each stage of a supply chain. These control strategies should be able to avoid the bullwhip effect.

This model is now being developed to simulate a seasonal ordering behavior. The model is designed in the following manner:

For the examination of the bullwhip effect a model of a supply chain according fig. 1 will be used. The behavior of each stage is the same. The time to place an order is 1 time unit. The time for delivery is 3 time units. The lead time to fill up the stock is for one stage 4 time units. If a customer places an order the lead time for the entire supply chain is 16 time units to deliver the material. To be able to deliver within the lead time of one stage, each stage needs a stock.

The only decision in this simulation is to decide about the quantity of the order to compensate a difference in the own inventory. All other influences are eliminated. The applied controlling strategies for this decision will be described in chap. 2. This decision has been taken each time unit. It is obvious that these parameters do not

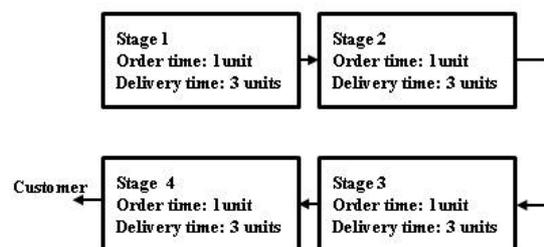


Figure 1: Model of a Supply chain

simulate a real supply chain. Normally the lead time is much shorter than the time for the next order. However,

this simulation demonstrates with this short order period the bullwhip effect in an impressive manner.

2 DYNAMIC BEHAVIOR OF A SUPPLY CHAIN

Before the dynamic behavior of a supply chain will be examined, a suitable closed-loop controller for a particular stage in the supply chain has to be found. Assuming the unrealistic precondition of a zero lead time the best strategy is: "input is output". Under this precondition there is no need for a stock at all. Now this strategy is applied to a supply chain as described above (fig. 2).

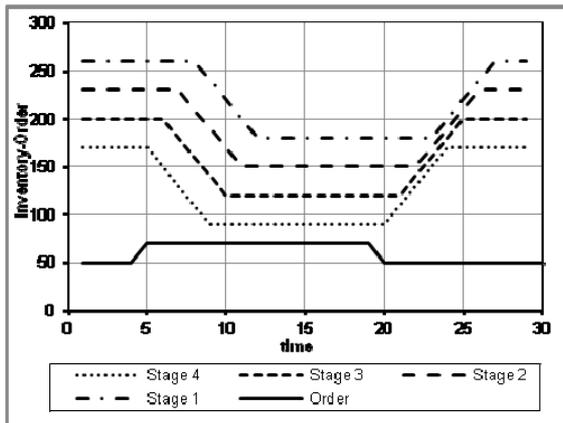


Figure 2: Stock with input=output strategy

If the customer increase his order, here from 50 to 70 items, the stock of stage 4 decrease in a linear manner. The other stages follow after the order time of 1 unit. After the lead time the stock is constant, because now the output of the stock is equivalent to the input. However there is a difference to the nominal stock. Does the customer reduce his order to the original value, the behavior is vice versa.

Assuming the increase or decrease in the order is permanent, the aim of each particular stage is to

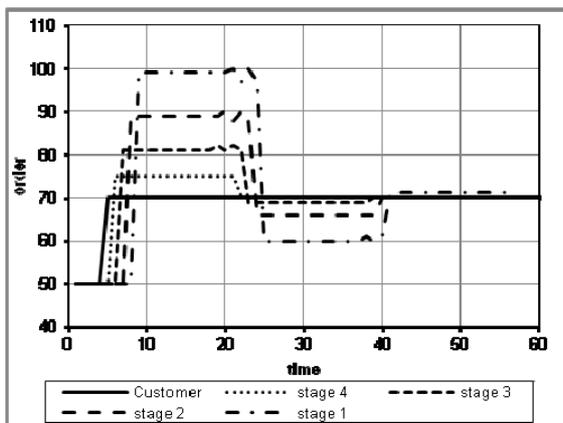


Figure 3: Orders with compensation strategy: constant order within 16 time units

equalize this difference to the nominal stock, which occurred with the strategy "input is output". Therefore

the orders have to be increased for a certain time (fig. 3). In this example the time for compensation is 16 time units. If the compensation time is constant for all stages, the stages upstream have to increase their orders more and more. The reason is that they have to compensate their own stock difference and additional the stock differences in the stages downstream.

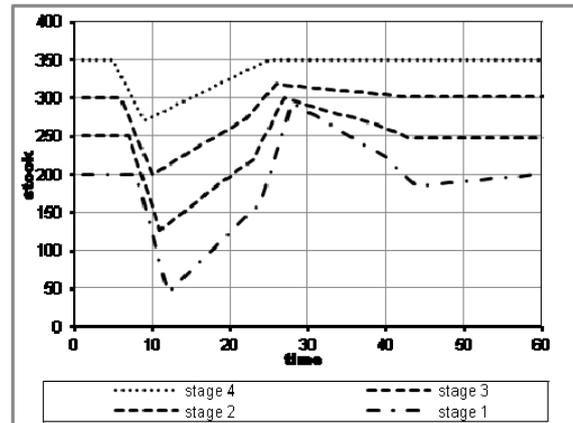


Figure 4: Stock with compensation strategy: constant order within 16 time units

Only the stage at the very end of the supply chain is able to compensate the stock difference within the scheduled time (fig.3 and fig. 4). For all other stages it requires more than double the time. This is quite obvious: The last stage has only to fulfill the customer's requirement. All other stages have to fulfill the customer's requirement and have to compensate the stock difference of all stages downstream. Only when the first stage in the supply chain has balanced the stock difference, the order is reduced to the value of the customer. This is the reason why the bullwhip effect also occurs in the stock (fig.4).

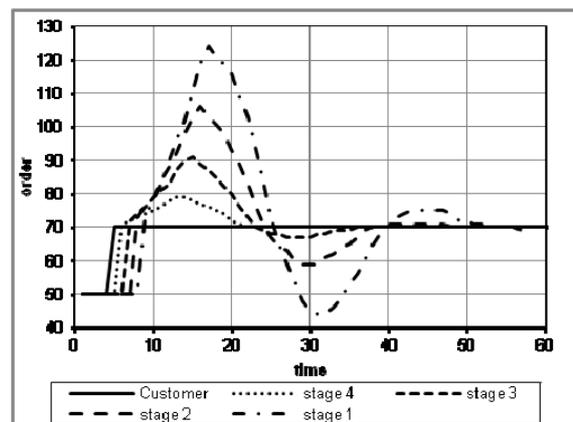


Figure 5: Order strategy of the closed-loop controller: linear increase and linear decrease within 16 time units

The second strategy is relative similar. Instead of a constant order over a certain time there is a linear

increase and a linear decrease of the orders (fig. 5). The result is a smoother behavior in the orders and the stock (fig.6).

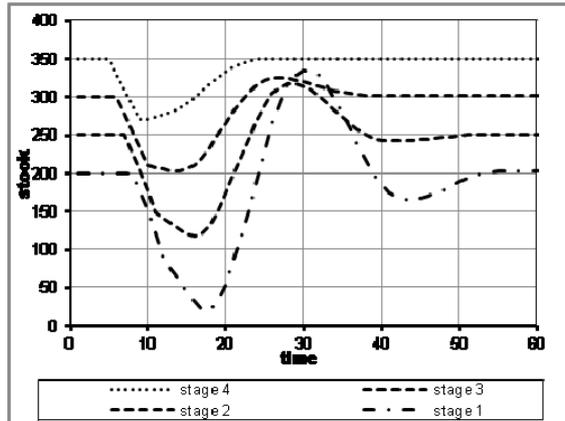


Figure 6: Stock with linear increase and linear decrease order strategy

For both strategies the quantity of these additional orders is calculated by multiplying the lead time and the difference in the customers order. Due to the increasing orders over time for the second strategy there is a higher peak in the orders. This leads to a larger bullwhip effect in the orders and the stock. The time for compensation is more or less the same for both strategies.

For the following examination it seems to be not so important which strategy is applied. For the study, the second strategy is chosen.

3 SEASONAL TREND

A seasonal trend with oscillating orders also leads to major changes in inventories. Therefore the aim must be to minimize the oscillation of the stock by appropriate closed-loop control. If the oscillation of the stock is minimized, then the average stock is at a minimum too. A seasonal trend is simulated by a sine function very well. In this simulation the amplitude of the sine is +/- 10% of the average. The following simulation examines the fluctuation of the stock for the individual stages in the supply chain (fig.7). The diagram shows the fluctuation of the stock between a maximum and minimum for all stages of the supply chain as a function of the period of the seasonal trend. For a very large period lengths there are only small fluctuations of the stock in all stages of the supply chain. The bullwhip effect is very small. For shorter period lengths the fluctuation in the stocks increases rapidly. The bullwhip effect at the beginning of the supply chain is tremendous.

There is a strong influence of the period length of the seasonal trend to the quality of the closed-loop controlled stocks of each stage. In general terms: The

longer the period length of the seasonal trend, the smaller the stock difference in each stage.

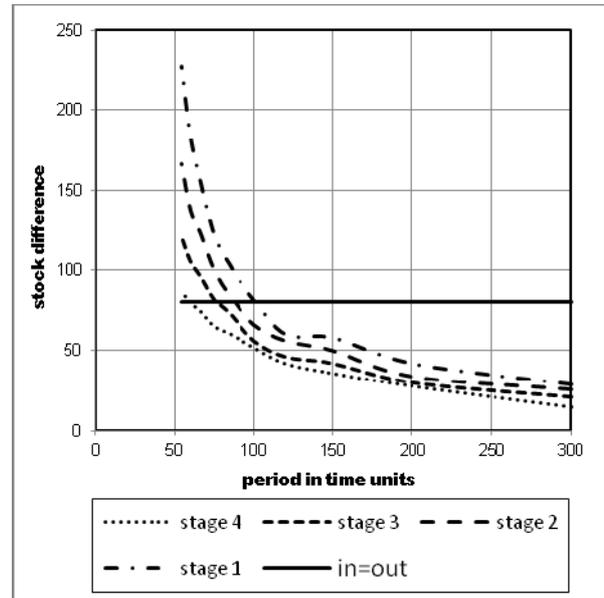


Figure 7: Stock differences for closed-loop controlled supply chain (compensation time 16 time units and in=out strategy)

The quality of this closed-loop controlled supply chain is now compared with two other strategies.

1. Input=output (fig.7 and fig.8)
2. Constant order (fig.8)

The controlling strategy input=output is a perfect strategy for systems, where the lead time is zero. In this model with a lead time of 16 time units the strategy gives a permanent deviation in the stock of all stages. Therefore, all curves are identical (fig. 7). The simulation has shown that regardless of the period always occurs a constant deviation of 80 units in the stock. This variation is also independent of the station because all stations always get the same order only with a defined time delay. Fig. 7 demonstrates that this strategy is better than the closed-loop control, if the period is lower than a definite value. This value depends from the stage in the supply chain. For stage 1 the strategy is better for period lengths lower than 100 and for stage 4 for period lengths lower than 70.

The reason for the bad results of the closed-loop control at short period lengths is the lead time. The control could not react fast enough to compensate the variation in the orders. The stages of the supply chain are acting like an oscillator.

The strategy constant order, which is the average of the sine, is not at all a closed-loop control. The variation in the orders has to be compensated with a tremendous high stock in the last stage of the supply chain. Theoretically all other stages upstream do not need any stock, because they get only constant orders from their

customers. All the variation in the stock takes place at the end of the supply chain.

In the next step these results are compared with the input=output strategy and the closed-loop control. This needs a bit different point of view. For this strategy the oscillation in the stock only take place in the stage close to the customer. Theoretically the other stages do not need any stock, because they get a constant order every time. For the comparison all stock differences are added (fig.8). This is now a view to the success of the entire supply chain. For long periods of the seasonal trend the closed-loop control gives the best results for the entire supply chain. If the period falls under a critical value, the strategy “ordering the average” gives better results. The strategy in=out does not give best results in any case.

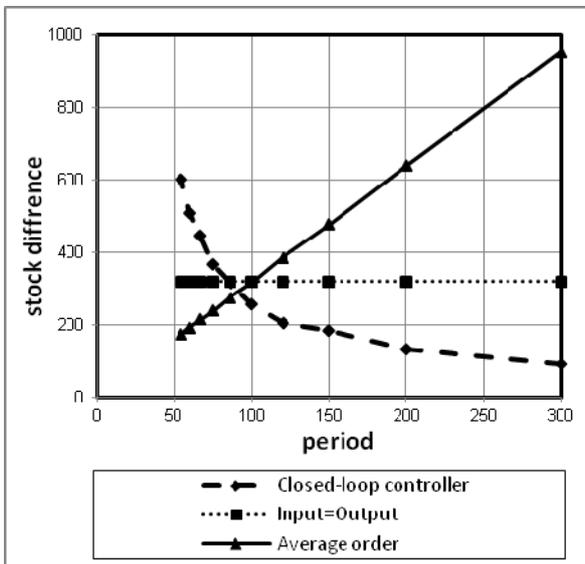


Figure 8: Comparison of the strategies for the entire supply chain

5 SUMMARY AND CONCLUSIONS

This study is a theoretical view of the dynamics in a supply chain. For this examination a quite simple model has been used. The advantage of a model like that is to see the main influences of the dynamic behavior of the supply chain.

The target of all stages is to keep the stock at a minimum with a seasonal trend of the customer's orders. This has been realized by a closed-loop control. The only decision which could be done was the quantity of the orders. Due to lead times caused by orders and delivery, it is difficult or better more or less impossible to get a constant stock by applying a closed-loop

control. The seasonal trend has a tremendous influence on the stock: The shorter the period the higher the oscillation in the stock.

For large periods each stage of the supply chain can control its stock independently and get good results. However an increasing bullwhip effect is visible for the stages downstream. Under a critical value of the period the strategies “input=output” and “average order” give better results than the closed-loop controller. Under this critical value the stages of the supply chain are acting like an oscillator. For the “average order strategy” a cooperation of the stages of a supply chain is necessary, because all the variation takes place in the stock close to the customer. It is questionable if such a cooperation can realized in reality. However, this strategy gives the best results for the entire supply chain.

Subsequently, it has to be checked if this theoretical knowledge can be transferred to a real supply chain.

REFERENCES

- Barbey, H.-P.: Dynamic Behaviour of Supply Chains. Proceedings of 27. European Conference on Modelling and Simulation (ECMS), Alesund, 2013, 748-752.
- Barbey, H.-P.: A New Method for Validation and Optimisation of Unstable Discrete Event Models“, appeared in proceedings of 23. European Modelling & Simulation Symposium (EMSS), Rome, 2011.
- Barbey, H.-P.: Simulation des Stabilitätsverhalten von Produktionssystemen am Beispiel einer lagerbestandsgeregelten Produktion, appeared in: Advances in Simulation for Production and Logistics Application, Hrsg.: Rabe, Markus, Stuttgart, Fraunhofer IRB Verlag, 2008, S.357-366.
- Barbey, H.-P.: Application of the Fourier Analysis for the Validation and Optimisation of Discrete Event Models appeared in proceedings of ASIM 2011, 21. Symposium Simulationstechnik, 7.9.-9.9.2011, Winterthur.
- Bretzke, W.-R.: Logistische Netzwerke, Springer Verlag Berlin Heidelberg, 2008.
- Dickmann, P.: Schlanker Materialfluss, Springer Verlag Berlin Heidelberg, 2007.
- Erlach, K.: Wertstromdesign, Springer Verlag Berlin Heidelberg, 2010.
- Forrester, J.W.: Industrial Dynamics: A major breakthrough for decision makers. In: Harvard business review, 36(4), 1958.
- Gudehus, T.: Logistik, Springer Verlag Berlin Heidelberg, 2005.

AUTHOR BIOGRAPHIES

HANS-PETER BARBEY was born in Kiel, Germany, and attended the University of Hannover, where he studied mechanical engineering and graduated in 1981. He earned his doctorate from the same university in

1987. Thereafter, he worked for 10 years for different plastic machinery and plastic processing companies before moving in 1997 to Bielefeld and joining the faculty of the University of Applied Sciences Bielefeld, where he teaches logistic, transportation technology, plant planning, and discrete simulation. His research is focused on the simulation of production processes.

His e-mail address is:

`hans-peter.barbey@fh-bielefeld.de`

And his Web-page can be found at

<http://www.fh-bielefeld.de/fb3/barbey>

THE TYCHE AND SAFE MODELS: COMPARING TWO MILITARY FORCE STRUCTURE ANALYSIS SIMULATIONS

Cheryl Eisler and Slawomir Wesolkowski
Centre for Operational Research and Analysis
Defence Research and Development Canada
101 Colonel By Drive, Ottawa, ON K1A 0K2 Canada
Email: cheryl.eisler@drdc-rddc.gc.ca
Email: s.wesolkowski@ieee.org

Daniel T. Wojtaszek
Atomic Energy of Canada Ltd.
Chalk River Laboratories
1 Plant Road, Chalk River, ON K0J 1J0 Canada
Email: wojtaszd@aecl.ca

KEYWORDS

Force Structure Analysis, Fleet mix analysis, Capability-Based Planning, Discrete Event Simulation, Multi-Objective Optimization, Scheduling.

ABSTRACT

In the past, several force structure analyses have been conducted for the Canadian Armed Forces using moderate fidelity (e.g., Tyche) and low-fidelity (e.g. Stochastic Fleet Estimation or SaFE) simulation models within optimization frameworks. Monte Carlo discrete event simulations like Tyche are computationally expensive and can only be used in optimizations that require few force structure evaluations. The SaFE model acts as a simple surrogate model that can be utilized by more global optimization techniques. SaFE, originally developed to study air mobility fleets, was adapted to accommodate a larger set of capabilities and more scheduling heuristics so that the performance of many force structures can be quickly assessed while minimizing a set of objectives. The amount of time required to find the SaFE optimal force structures is significantly less than using Tyche. This indicates that SaFE could be an important tool for discovering pareto-optimal force structures (within the space of all possible mixes) that would represent practical lower bounds on the force structure requirements for accomplishing expected future scenarios. The purpose of this paper is to compare and contrast the use of Tyche and SaFE through simulation optimizations on a given dataset.

INTRODUCTION

Determining the best future military force structure, comprised of a set of assets, to accomplish a set of defence and security tasks is a challenging undertaking. The set of tasks must be thoroughly investigated; requirements, frequencies, and durations for each task require definition. Potential assets must be identified and their abilities to meet task requirements assessed. Besides the necessity for accurate data from which to model, the force structure problem is further complicated by the deep uncertainty (Bui et al. 2009) inherent in modelling future environments. Thus, a force structure must be capable of addressing many possible combinations of future operational tasks. Furthermore,

assets are large capital investments; accordingly, the goal is not only to find the appropriate force structure size and mix with respect to the devised future scenarios, but also the most capable structure at the lowest cost (Wojtaszek and Wesolkowski 2012).

Since large capital procurement projects undergo significant internal and external scrutiny, it is incumbent upon decision-makers to balance many conflicting objectives, justifying investments with anticipated needs. Due to the non-linear nature of the performance objective functions, as well as the length of computational time required to evaluate individual force structures, it is often not realistic to find a globally optimal structure in the time normally given to complete such studies. The computational complexity is exacerbated when searching for the pareto-optimal set of structures with respect to multiple objectives (Wojtaszek and Wesolkowski 2013). It is, therefore, critical that methodologies for quickly identifying optimal future force structures be investigated.

Two optimization-simulation approaches to force structure analysis used within the Defence Research and Development Canada's Centre for Operational Research and Analysis (DRDC CORA) are examined. The first approach uses a computationally intensive, Monte Carlo discrete event simulation model known as Tyche (Eisler and Allen 2012) within a direct search optimization framework. The model takes a top-down approach to test force structures, mimicking the decision of a military scheduler by assigning assets within a given force structure to scenarios as they arise. A single simulation run often requires hours to complete, and an optimization search can take weeks or months on today's desktop computers; necessitating an optimization procedure that requires relatively few steps to converge to the optimal force structure composition.

An alternative to the moderate-fidelity approach based on Tyche is the low-fidelity approach of DRDC CORA's Stochastic Fleet Estimation (SaFE) model (Wojtaszek and Wesolkowski 2013). SaFE is also a Monte-Carlo based simulation, which generates average yearly requirements from a dataset with frequency, duration, and capacity requirements for tasks (scenarios

without a stochastic location element) and assets. However, the total force structure requirements are estimated from the bottom up, through a fixed matching of assets to scenarios, and no attempt is made to account for scheduling constraints (e.g., start and end dates). Given SaFE's relatively quick run time (approximately one millisecond on the same data run through Tyche), optimization is carried out over the solution space of all possible task to asset assignments, not just all of the force structure compositions.

Both models will be described subsequently in further detail. The optimization results of the Tyche and SaFE models will be compared, and their roles for military force structure analysis contrasted.

THE TYCHE MODEL

Tyche schedules the deployment of assets within a force structure to address a set of missions (Eisler and Allen 2012). Figure 1 illustrates the implementation of the Tyche model. On the top right, a fixed set of demands is created: missions to which a military force structure should endeavour to respond. These missions are created as scenarios, and may be broken down into one or more phases. Each phase may be random or scheduled, with its own frequency, duration (and associated probability distribution) and possible theatre locations, as well as a set of capability demands.

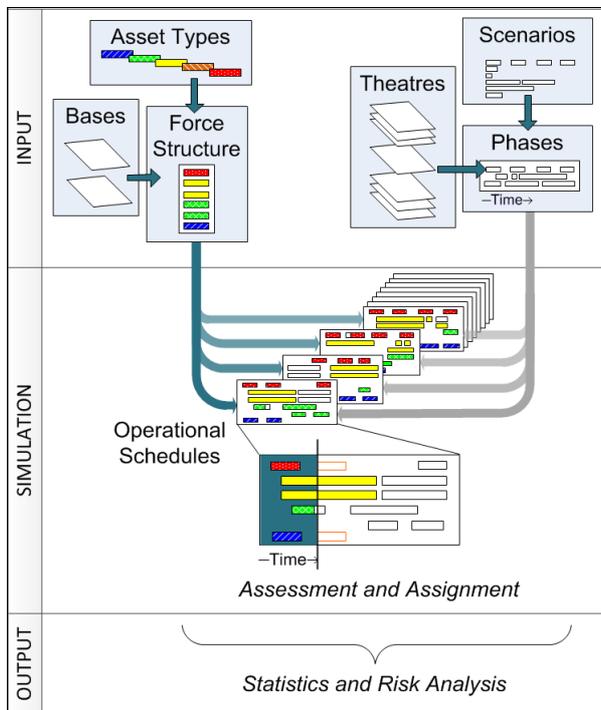


Figure 1: Tyche Model

Tyche can model a number of asset types, each supplying different capabilities. Force structures are constructed out of these asset types by specifying a quantity for each type and a physical location where

they are based. To run a Tyche simulation, one force structure is selected to test a capability supply from the set of assets against demand requested from the given scenarios. Demand is constructed stochastically from the scenarios for frequency, start date, and duration in the schedules. Scenarios can be randomly generated using a Poisson process or scheduled at known intervals; durations are generated using uniform or triangular distributions. Assets within a force structure are then assigned to the schedule chronologically utilizing the policy to meet a single requirement by selecting from a list of available assets based on information that is known and actionable at the moment a mission occurs (Wu et al. 2009). The available assets within the force structure are assessed a numerical score for the capability used in the scenario and optional penalties for excess capability supply, timeliness into theatre, and scheduling conflicts. The scoring algorithm (Eisler and Allen 2012) factors in the quality, quantity, and subjective weighting of importance of capabilities matched between the supply and demand for a specific combination of assets. The combination of assets with the highest score is then assigned to the scenario in the operational schedule.

This process is repeated for all simulation iterations in a Monte Carlo approach (Robert and Casella 2004), and force structure performance is evaluated based on how well and how often the scenario's capability requirements are met. This is done in the form of statistics gathered from the collection of operational schedules on unmet capability demand per scenario, and by factoring in the frequency of scenario occurrence and political impact of failure to meet such requirements, to form a metric of political risk.

Performance Metric

The average yearly political risk R for a set of scenarios is defined as

$$R = \sum_{\forall s} f_s I_s P_s \quad (1)$$

where the risk for a given scenario s is defined as the product of the annual frequency of occurrence f_s , the political impact I_s of scenario failure, and the percentage of time the capability supply deployed by the scheduler is inadequate P_s . The first factor is assessed by averaging the number of times the scenario occurs yearly across all schedules. The second factor, impact score, is provided by subject matter experts (SME) into the calculation. Each scenario is assigned to an impact category with an associated impact score. The third factor in the risk calculation is defined as a weighted calculation of the percentage of time that capability requirements are not met at various levels for the scenario (Eisler and Allen 2012).

Optimization Framework

Tyche was designed as a tool to evaluate and compare individual force structures. There is no optimization built in to drive the search for better force structure compositions. However, Tyche can be used inside an optimization framework, provided that the algorithm does not require a significant number of force structure evaluations due to the computational cost associated with each simulation run. An optimization is conducted within the solution space of all possible force structure compositions, with Tyche evaluating the performance of each feasible structure.

A force structure analysis study conducted internally by DRDC CORA used the Hooke-Jeeves algorithm (Hooke and Jeeves 1961), modified to combine a local exploratory search with a global pattern search, to perform the optimization procedure. Starting from an initial force structure composition, the exploratory search makes cumulative incremental changes to each asset type to determine if the objective value improved. The best combination of local improvements is used to drive the pattern search for larger step sizes. Although this algorithm can easily get trapped in local optima, it has two major advantages for application with Tyche. First, it requires few function evaluations, which are computationally costly. Second, it is simple enough not to require automation, given that manual input is required to set up force structures within Tyche.

Two primary objectives were defined to determine optimal force structures: minimizing total force structure risk and size. Due to the discrete political impact categories, the risk minimization was then defined in two ways, each used to drive separate optimizations. The first optimization minimized total risk and structure size, where a change in force structure was retained if the total risk decrease was deemed statistically significant. Noting that the standard error SE was estimated using the sample variance of the risk distribution divided by the square root of the number of schedule realizations and, assuming that the distribution can be normally approximated, the statistical significance was calculated in pairwise comparisons where the $\pm 2 SE$ intervals did not overlap (Payton et al. 2003).

A second optimization minimized risk per impact category until a threshold of acceptable risk (as defined by military SMEs) was reached. That is, a change in force structure was retained if the risk in any impact categories showed a statistically significant decrease. Again, the number of assets in the force structure was minimized by rejecting changes (i.e., with asset additions) that showed no statistically significant improvements. The search was terminated once the risk in each impact category met the given threshold within the bounds of the statistical significance. The procedure is illustrated by the following pseudo code on the force

structure of composition \bar{x} , a vector count of each asset type, α as the pattern search acceleration factor and $\bar{\delta}$ as the pattern search step vector:

```
procedure modifiedHJ( $\bar{x}$ ,  $\alpha$ ,  $\bar{\delta}$ ) with
  DO WHILE termination criterion not met
    // Exploratory search
     $\bar{\Delta}$  = step size, initially as vector of zeros for
    total number of asset types
    FOR  $i = 1$  to number of asset types
       $x_{i_{new}} = x_i + \delta_i$ 

      Evaluate simulation at  $\bar{x}_{new}$ 
      IF  $[R(\bar{x}_{new}) - 2SE_{R(\bar{x}_{new})}] < [R(\bar{x}) - 2SE_{R(\bar{x})}]$ 
         $\Delta_i = \delta_i$ 
      ENDIF
    NEXT
    // Pattern search
    DO WHILE  $\bar{x} \neq \bar{x}_{new}$ 
       $\bar{x}_{new} = \bar{x} + \bar{\Delta}$ 

      Evaluate simulation at  $\bar{x}_{new}$ 
      IF  $[R(\bar{x}_{new}) - 2SE_{R(\bar{x}_{new})}] \geq [R(\bar{x}) - 2SE_{R(\bar{x})}]$ 
         $\alpha = \alpha - 0.5$ , as acceleration factor
         $\bar{\Delta}_{new} = \alpha \bar{\Delta}$ , where all  $\Delta_i$  values must be
        integers and  $MIN(\Delta_i) = 1$ 
      ENDIF
    LOOP
  LOOP
```

A subsequent a greedy search is then applied to trim the solution. Trimming is carried out only if the modified Hooke-Jeeves algorithm is successful at either minimising the total risk to zero or the risk per category under the specified threshold. This trimming step is necessary since the pattern search can add several assets from different types at the same time, leading to a larger structure than necessary to achieve the specified objective. The trim procedure is illustrated by the following pseudo code on the force structure of composition \bar{x} :

```
procedure Trim( $\bar{x}$ )
  DO WHILE termination criterion met
    FOR  $i = 1$  to number of asset types
       $x_{i_{new}} = x_i - 1$ 

      Evaluate simulation at  $\bar{x}_{new}$ 
      IF termination criterion NOT met
         $x_{i_{new}} = x_i$ 
      ENDIF
    NEXT
    Evaluate simulation at  $\bar{x}_{new}$ 
```

DO WHILE termination criterion NOT met
 FOR all reductions to \bar{x}
 Select i with the maximum reduction of
 $R(\bar{x}_{new})$ and reverse change by
 $x_{i_{new}} = x_i + 1$
 NEXT
 Evaluate simulation at \bar{x}_{new}
 LOOP
 LOOP

The results of the optimizations of the Tyche runs will be discussed in comparison with the results of the SaFE simulation optimization (as conducted on the same input data set) after a description of the SaFE model and its optimization framework is given.

THE SAFE MODEL

Like Tyche, SaFE is also a capability-based model that uses a Monte-Carlo approach to determine possible force structures based on the tasks that must be performed. It uses a dataset of task frequency, asset- and task-specific durations, and capability (in the case of air mobility (Wojtaszek and Wesolkowski 2013), these were passenger and freight capacities) requirements to derive demand over a stochastically generated number of tasks. The force structure is built from the bottom up, where its composition is computed such that there are sufficient assets to accomplish an average set of tasks. Since assets are matched to tasks via capabilities, there can be many assignment combinations. Force structures generated by SaFE are input into an (usually multi-objective) optimization procedure so that assets can be traded off against each other based on common capability. Given that SaFE is a bottom-up task-driven model, if in one solution the number of assets of a particular type increases (in comparison to another solution), then the number of assets of a different type which has similar capability will usually decrease.

To illustrate the differences between SaFE and Tyche, consider Figure 2. Instead of building a force structure out of a variety of asset types at a number of bases to test during a simulation, SaFE exhaustively matches each task to a specific asset or groups of assets. This asset to task assignment is done in a capability-based manner ahead of the optimization proper in order to limit the solution space to all feasible asset assignments. Each individual asset to task assignment is known as a configuration.

On the top right of Figure 2, demand is generated stochastically from a set of tasks, using frequencies and durations derived from triangular distributions. Asset-specific duration distributions (uniform) are also defined for completion of each task, and computed based on the configuration in use. For each iteration, the total demand can be calculated as time required for each asset type.

The total time for each asset type is then averaged over all iterations to form the average annual demand. The number of assets required in the force structure to satisfy this average level of demand is computed simply as the whole number of assets that can provide such time (for example, 2.6 years of average annual demand requires 3 assets within the force structure). The sample variance of these durations is also computed to determine how much the demand varies across all of the scenarios.

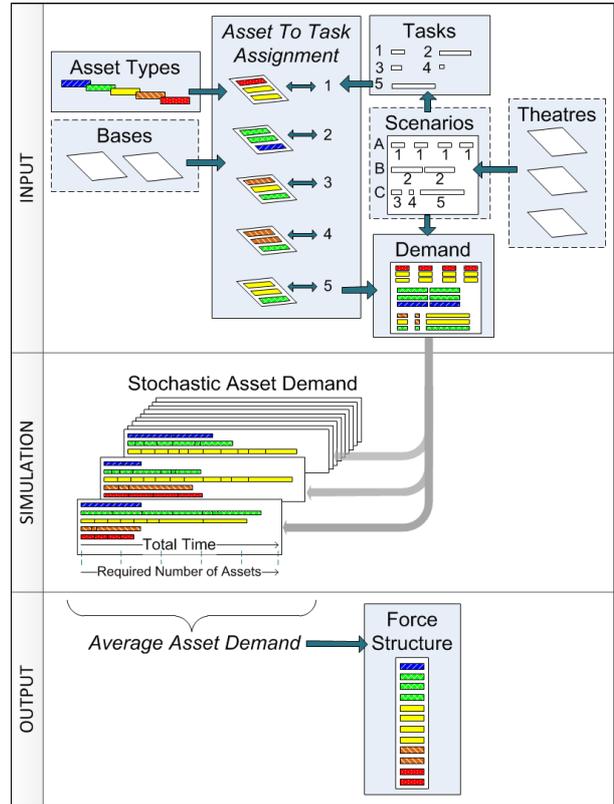


Figure 2: SaFE Model

Essentially, SaFE assumes a much simplified world where only total time on task for each asset type is needed to compute force structure requirements. It does not take into consideration event scheduling, such as task start and end times, task cancellation or prioritization, or other assignment constraints. SaFE yields the best possible representation of task requirements and, thus, underestimates realistic task requirements to produce a lower bound on a required force structure. Due to this simplification, SaFE can be used in an optimization framework to generate and evaluate force structures much more quickly than even the most efficient Monte Carlo discrete event simulation. This improved speed is vital when searching for optimal structures, a process that requires many force structures to be evaluated.

For analyses where higher fidelity is required, the SaFE model could be exploited as a preprocessing tool. It may reduce the problem space by eliminating large number

of inefficient options and, thus, reduce the cost of using higher fidelity tools such as Tyche.

Adaptation to New Data

In Figure 2, there are several objects indicated with dashed lines, such as bases, theatres, and scenarios. These are common concepts between Tyche and SaFE; however, SaFE does not provide direct support for such data entry. To accommodate these concepts, the following adaptations were made:

- Multiple bases: assets of the same type defined at different locations and with different transit times to theatre (as asset-specific task completion durations).
- Scenarios with a probability of occurring at more than one theatre and/or more than one phase per scenario: handled through data manipulation to obtain a suitable equivalent of multiple tasks in SaFE.
- Asset assignment dependent upon availability at the time a scenario arises in the simulation: Tyche would send the same unique set of assets to a scenario every time if there were no limits on the number of assets available, but to enable the SaFE model to use an optimization mechanism for configuration generation, possible asset to task assignments are calculated as those that provide all the capabilities at the required level while also providing a minimum of excess capability.
- The effect of asset types that act as force multipliers: captured by modelling a single additional asset type with enhanced capabilities.

Performance Metric

The objective of the optimization is to search for force structures that are capable of fulfilling the average requirements and are minimal with respect to size, scenario duration, and risk of failure.

The force structure size objective (E_{size}) is an evaluation of the number of assets resulting from the chosen configuration and identifies structures which require minimal resources but are still capable of accomplishing the average scenario. The size objective is defined as

$$E_{size} = \sum_{a \in m} F(a) + w \cdot F(m) \quad (3)$$

the summation of the number (F) in each asset type (a) plus a small weighted ($w=0.01$) total to account for the number of a single type of relatively low-value force multiplier assets (m).

The scenario duration objective function evaluates the average time it takes to accomplish a scenario. The duration objective (E_{time}) is defined as

$$E_{time} = \sum_s \delta(s), \text{ where } \delta(s) = \max_a (d_s(a)) \quad (4)$$

where $d_s(a)$ is the time it takes one asset of type a to accomplish its portion of all instances of scenario s , and $\delta(s)$ is the maximum time it would take any of the assigned assets to complete the scenario (thus the duration of the asset that travels furthest to the theatre is the one that defines the duration for the whole configuration). This assumes that all assets travel at the same speed, and that all assets must arrive at the theatre before the scenario can commence.

The force structure size and scenario duration objectives are evaluated using the average duration output from SaFE. However, the requirements of any iteration may vary from that of the average iteration. To mitigate the effects of this uncertainty, a risk-based objective is used, which is an evaluation of the ability of a configuration to accomplish all iterations. The risk objective (E_{risk}) is computed by

$$E_{risk} = 1 - \prod_a \pi(a) \quad (5)$$

as the probability that at least one asset will not be able to accomplish its requirements. The probability that an asset will be able to fulfill its requirements is given as $\pi(a)$ (Willick et al. 2010).

Optimization Framework

A single simulation run in SaFE is conducted for a given asset to task assignment configuration over 10^4 iterations (typically) of one year in duration each. An average force structure can then be calculated to meet the average set of demand over all iterations. The space of all possible configurations is very large (Wojtaszek and Wesolkowski 2013) – significantly larger than the force structure composition solution space. Since this large configuration space cannot be exhaustively searched in a practical amount of time, a metaheuristic is required.

Given that multiple objectives are considered, a multi-objective optimization algorithm needs to be used to provide a set of non-dominated solutions with respect to these objectives. Among the multi-objective algorithms that exist (Deb 2005), a well-studied one that has been utilized previously with SaFe is the Non-Dominated Sorting Genetic Algorithm-II (NSGA-II). NSGA-II is an elitist evolutionary algorithm that groups individual solutions into non-dominated fronts, and uses a crowding-distance operator to preserve diversity of solutions (Deb et al. 2002). Each solution comprises a configuration of asset to task assignments, and a base distribution for each asset. The NSGA-II pseudo code is not provided here, as it is adequately given in a variety of references, including (Deb et al. 2002).

RESULTS

The study dataset included 164 scenarios and 28 theatres. There were 14 asset types modeled, each at two

possible bases. The results of the asset to task assignment algorithm generates 7.4×10^{62} possible configurations over all scenarios.

The NSGA II was run 50 times with 1 000 individuals (configurations) for 10 000 generations each with a mutation rate of 20%. Multiple runs were used to ensure the repeatability of the results obtained with respect to quality. The quality of the results was assessed using a hyper-volume measure (Fleischer 2003). The non-dominated fronts of the last generation over each run were combined into a single set of individuals, and then the non-dominated sorting algorithm was performed on this set to give the combined non-dominated front over the solutions from the 50 runs. The hyper-volume of the last generation of each run was then computed and compared to the hyper-volume of the combined non-dominated front. The hyper-volume average and standard deviation over all the runs corresponded to $96\% \pm 3\%$ with respect to the combined best non-dominated front. Therefore, the quality of the results from each run can be considered to be similar to the others, and, therefore, analysis in the remainder of this section is carried out on the results of one of the runs.

Figure 3 shows a plot of E_{time} versus E_{size} for the 81 configurations in the non-dominated front, with the colour of each point representing the value of E_{risk} . This figure shows the trade-off between the size of the force structure and the risk of not being able to fulfill all of the demand in an iteration.

When looking at configurations with the same value of E_{size} , configurations with lower E_{time} have higher E_{risk} , thus demonstrating that there is a risk of not being able to assign assets from the closest base to theatre. The lowest value of E_{risk} over the non-dominated configurations is 0.27, indicating that the duration of asset use in an iteration may deviate significantly from the average. Recall that E_{risk} does not take into account the timing of scenarios within an iteration and the requirement that scenarios must be performed within time windows. Therefore, the risk of a force structure produced by optimizing SaFE not being able to satisfy all of the demand in a given iteration may be greater than E_{risk} .

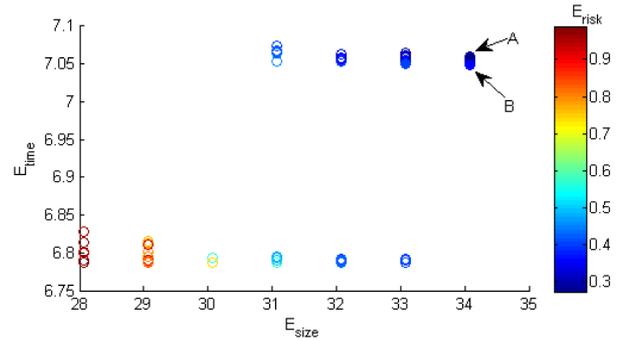


Figure 3: Plot of Objective Values for Configurations in the Non-Dominated Front

As mentioned previously, the same force structure can be computed from different configurations. For example, configurations A and B shown in Figure 3 both result in the same force structure, but configuration A has lower E_{risk} (0.27 vs. 0.36) and higher E_{time} (7.06 vs. 7.05). These differences are due to differences in the assets assigned to each scenario and the base from which the assets are assigned.

Within the 81 non-dominated configurations, there are 24 distinct force structures. The ranges of number of each asset type over these structures are shown in Table 1. When compared to the force structures run through Tyche (three distinct force structures were used to seed the initial values for two separate optimizations to produce six final structures), the upper bounds on these ranges are similar to or slightly less than the SME force structures for most assets. This result indicates that the SME force structures could theoretically satisfy the average iteration requirements with respect to the duration of asset usage with some unused asset capacity. The force structures from the optimization conducted using Tyche are much larger in number for most asset types than the upper bound on the range of non-dominated structures, indicating that they could theoretically satisfy the average iteration requirements with a large amount of unused asset capacity (which may be necessary to meet scheduling constraints).

Comparison of Suggested Force Structures

The force structure compositions produced using the SaFE model were run back through the Tyche

Table 1: Range of Number of Assets in Recommended Force Structures

	Type 1		Type 2		Type 3		Type 4		Type 5		Type 6		Type 7		Type 8		
	Base	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
SaFE	3-4	2-3	2	2	1-2	2	2-4	4-7	0-1	1	1	1	1-2	3-5	2-3	4-5	
SME	3-6	3-6	2	2	2	2	5-6	6-7	1	2	0-1	0-1	3	5	7	8	
Tyche	10-14	10-14	5-6	5-6	3-4	4	10-14	11-14	1-2	2-3	1-2	2-3	1-3	1-5	6-8	7-8	

simulation in order to compare results with common metrics. Each force structure was run for 1 000 iterations. Of the three SaFE objectives, only E_{size} , a good indicator of force structure size, is independent of the model. E_{time} and E_{risk} are associated with specific asset to scenario assignment configurations, of which there may be multiple for the same force structure composition, and cannot readily be generated for the SME or Tyche recommended force structures. As a result, comparisons will primarily be made on correlations between E_{size} , political risk, and E_{risk} . The set of force structures chosen for this comparison comprise all of the force structures from the final generation of the NSGA-II, not just the 24 in the non-dominated front. This set comprises 274 distinct structures, and was chosen to provide a better statistical analysis. In the amount of time it took to run an evolutionary optimization procedure to find 24 non-dominated force structures (less than 24 hours), the Tyche simulation was only able to evaluate approximately 77 force structures (2.5 hours per force structure, running 8 simulations in parallel).

Performance evaluations using SaFE are not as precise when compared to Tyche because the SaFE evaluations are based on average requirements. In addition, the risk measures used are not directly comparable, since the political risk objective is a weighted sum of stochastic scenario performance, where each scenario is weighed according to the political impact of not being able to provide its required capability. The E_{risk} objective, on the other hand, does not distinguish between the importance of different scenarios.

Another issue with comparing E_{risk} and the political risk for a force structure is that there may be multiple values of E_{risk} for a given structure due to the possibility of multiple configurations for the asset to task assignment. In order to determine which value of E_{risk} to use for each force structure, the correlation coefficient is computed between the structure's political risk and each of the minimum, mean, and maximum values of E_{risk} . The resulting correlation coefficient values are 0.62, 0.63, and 0.63, respectively; thus indicating that there is very little difference among these values. All that can be said here is that the higher values of E_{risk} for a force structure may be slightly more reflective of the political risk computed using Tyche than the lower values. The mean value of E_{risk} will be used for the remainder of this section with the assumption that using either of the other values will not significantly change the analysis. The positive correlation obtained here shows that there is some potential in using SaFE to estimate the risk of a force structure, although Figure 4 shows that there are force structures with lower total political risk but larger E_{risk} than other structures; therefore, more work would be required to formulate a risk measure usable with SaFE that is more reflective of the political risk measure.

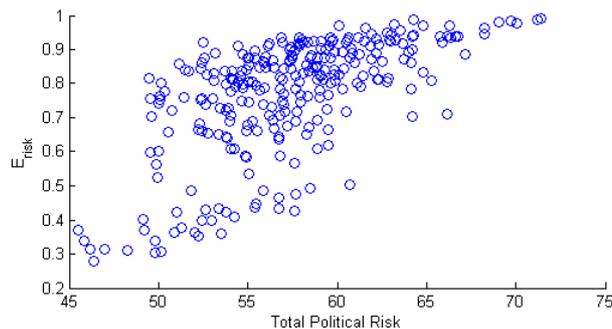


Figure 4: E_{risk} vs. Total Political Risk for the SaFE-produced Force Structures

By plotting E_{size} versus total political risk for the 274 SaFE, 2 SME, and 6 Tyche-recommended force structures, Figure 5 is obtained. There are three distinct clusters in the graph: the low political risk structures recommended by the Tyche optimization, the smaller SME-recommended structures with higher risk, and the even smaller SaFE generated structures with yet higher risk. From this plot, it can be seen that SaFE-recommended structures have the highest political risk and lowest size, while Tyche-recommended structures have the lowest political risk and the largest size.

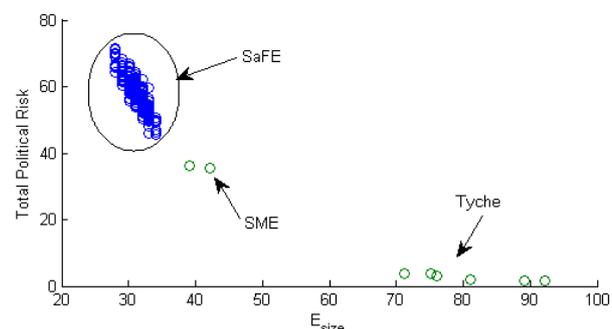


Figure 5: Size vs. Total Political Risk Objectives for All Recommended Force Structures

In SaFE, the assumption is that the occurrence of all tasks can be arranged in the most advantageous way for the entire force structure over time. It is clear that computing a force structure using a model based on several problem simplifications such as SaFE results in structures of lower size and, as a consequence, higher political risk. Figure 5 shows that SaFE and Tyche could be used as lower and upper bounds, respectively, on the number of assets needed within a force structure, and thus provides decision makers with realistic force structure size bounds.

CONCLUSIONS

The SaFE model was successfully used in a multi-objective optimization framework to find optimal force structures with objectives to minimize fleet size,

scenario duration, and risk of failure. A set of SaFE-derived force structures was then evaluated using the Tyche simulator in order to assess their political risk. The results showed that SaFE-recommended force structures have the highest political risk and lowest size, while the Tyche-recommended force structures have the lowest political risk and the largest size. Thus, results from SaFE and Tyche could be used respectively as lower and upper bounds on the number of assets required within a force structure, and provide decision makers with more realistic bounds on the political risk objective. SaFE appears to provide a lower bound on the force structure size since it is a model based on several constraint relaxations. In addition, there is some correlation between total political risk and E_{risk} although E_{risk} was not designed to estimate political risk.

The amount of time required to find the SaFE non-dominated configurations was less than 24 hours, whereas running a Tyche simulation required 2.5 hours per force structure; therefore, SaFE should be investigated further as a quick preprocessing tool that can sort through vast numbers of structures which can then be analyzed in more detail in Tyche. Furthermore, SaFE can also be modified to compute force structures that are capable of satisfying different levels of iteration requirements. For example, instead of using mean asset durations, the asset durations could be chosen such that they are greater than those of a user-specified percentage of iterations.

ACKNOWLEDGEMENTS

Dr. Wojtaszek's contribution to the work reported on in this publication was performed while he was employed at DRDC CORA. The authors would like to thank Leanne Stuiwe, a co-operative education student from the University of Waterloo, who assisted in the initial application of SaFE to the dataset discussed herein.

REFERENCES

- Bui, L.T.; M. Barlow; and H. Abbass. 2009. "A Multi-Objective Risk-Based Framework for Mission Capability Planning." *New Mathematics and Natural Computation (NMNC)*, No.5 (2), 459-485.
- Deb, K. 2005. "Multi-Objective Optimization." In *Search Methodologies*, Burke, E.K. and G. Kendall (Eds.). USA, Springer, 273-316.
- Deb, K.; A. Pratab; S. Agarwal; and T. Meyarivan. 2002. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II." *IEEE Transactions on Evolutionary Computation*, No.6 (2), 182-197.
- Eisler, C. and D. Allen. 2012. "A Strategic Simulation Tool for Capability-Based Joint Force Structure Analysis." In *Proceedings of the International Conference on Operations Research and Enterprise Systems* (Vilamoura, Portugal, Feb. 4-6) INSTICC, 21-30.
- Fleischer, M. 2003. "The Measure of Pareto Optima: Applications to Multi-Objective Metaheuristics." In *Evolutionary Multi-Criterion Optimization*, Fonseca, C.; P. Fleming; E. Zitzler; L. Thiele; and K. Deb (Eds.). Springer Berlin / Heidelberg, 519-533.
- Hooke, R. and T.A. Jeeves. 1961. "Direct Search" Solution of Numerical and Statistical Problems." *J. Assoc. for Computing Machinery*, No.8 (2), 212-229.
- Payton, M.E.; M.H. Greenstone; and N. Schenker. 2003. "Overlapping Confidence Intervals or Standard Error Intervals: What Do They Mean in Terms of Statistical Significance?" *Journal of Insect Science*, No.3 (34), 1-6.
- Robert, C.P. and G. Casella. 2004. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, NY.
- Willick, K.; S. Wesolkowski; and M. Mazurek. 2010. "Multiobjective Evolutionary Algorithm with Risk Minimization Applied to a Fleet Mix Problem." In *Proceedings of the IEEE Congress on Evolutionary Computation* (Barcelona, Spain, Jul. 18-23).
- Wojtaszek, D. and S. Wesolkowski. 2012. "Military Fleet Mix Computation and Analysis." *IEEE Computational Intelligence Magazine*, (Aug), 53-61.
- Wojtaszek, D. and S. Wesolkowski. 2013. "Evaluating the Flexibility of Military Air Mobility Fleets." *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, No.44 (4), 435-445.
- Wu, T.T.; W.B. Powell; and A. Whisman. 2009. "The Optimizing-Simulator: An Illustration Using the Military Airlift Problem." *ACM Transactions on Modeling and Computer Simulation*, No.19 (3), 1-31.

AUTHOR BIOGRAPHIES

MS. CHERYL EISLER obtained her M.A.Sc. from Carleton University in aerospace engineering. She works for DRDC CORA, where she leads research in the field of simulation for force structure analysis.

DR. SLAWOMIR WESOLKOWSKI is a scientist with DRDC CORA. He is also an Adjunct Professor with the University of Waterloo, where he obtained his Ph.D. in systems design engineering. He is interested in operations research problems and risk analysis.

DR. DANIEL WOJTASZEK received a Ph.D. degree in electrical engineering and joined DRDC CORA for two years as Post-Doctoral fellow, before taking a full time position as an Operations Research Analyst with Atomic Energy of Canada Ltd.

COPYRIGHT NOTICE

The authors of this paper (hereinafter "the Work") carried out research on behalf of Her Majesty the Queen in right of Canada. Despite any statements to the contrary in the conference proceedings, the copyright for the Work belongs to the Crown. ECMS 2014 was granted a non-exclusive license to translate and reproduce this Work. Further reproduction without written consent is not permitted.

MODELING OPTIMAL ALLOCATION CENTERS IN GIS BY FUZZY BASE SET OF FUZZY INTERVAL GRAPH

Leonid S. Bershtein
Scientific and Technical Center
"Information Technologies"
Southern Federal University
347900, Taganrog, Russia
E-mail: lsbershtyayn@sfedu.ru

Alexander V. Bozhenyuk
Scientific and Technical Center
"Information Technologies"
Southern Federal University
347900, Taganrog, Russia
E-mail: avb002@yandex.ru

Stanislav L. Belyakov
Scientific and Technical Center
"Information Technologies"
Southern Federal University
347900, Taganrog, Russia
E-mail: beliacov@yandex.ru

Igor N. Rozenberg
Public Corporation "Research and
Development Institute of Railway
Engineers"
Department of Computer Science
109029, Moscow, Russia
E-mail: I.rozenberg@gismps.ru

KEYWORDS

Fuzzy interval, fuzzy graph, fuzzy interval graph, syntactically-dependent linguistic variable, fuzzy set of interval bases.

ABSTRACT

In this paper the problem of optimal location of service centers is considered by minimax criterion. It is supposed that the information received from GIS is presented like graph with fuzzy intervals. The notion of fuzzy set of interval bases is considered. It is shown that the problem of service centers location is reduced to a problem of finding fuzzy set of interval bases. Method of finding fuzzy interval bases which is generalization of Maghout method for fuzzy graphs is suggested to use. In addition method of calculation of membership function of fuzzy intervals is proposed.

INTRODUCTION

Large-scale increasing and versatile introduction of geographical information system (GIS) is substantially connected with necessity of perfection of the information systems providing decision-making. GIS are applied practically in all spheres of human activity. Geographical information technologies have now reached an unprecedented position, offering a wide range of very powerful functions such as information retrieval and display, analytical tools, and decision support (Clarke and Englewood 1995; Longley et al. 2001).

Unfortunately, geographical data are often analyzed and communicated amid largely non-negligible uncertainty. Uncertainty exists in the whole process from geographical abstraction, data acquisition, and geoprocessing to the use (Zhang and Goodchild 2002).

The processes of abstracting and generalizing real forms of geographical variation in order to express them in a discrete digital store are defined as data modeling (Goodchild 1989), and this process produced a conceptual model of the real world. It is highly unlikely that geographical complexity can be reduced to models with perfect accuracy. Thus the inevitable between the modeled and real worlds constitutes inaccuracy and uncertainty, which may turn spatial decision-making sour.

One of the tasks solved with GIS is the task of the centers allocation (Kaufmann 1977). A search of optimum placing of hospitals, police stations, fire brigades and many other things the extremely necessary enterprises and services on some sites of considered territory are reduced to this task. In some cases the criterion of an optimality can consist in minimization of time of journey (or in minimization of distances) from the centre of service to the most remote service station. In other words a problem is optimization of "the worst variant" (Christofides 1976). However, very often, the information represented in GIS, contains approximate value or insufficiently authentic (Malczewski 1999). Therefore, at formalization of a problem of the centers allocation there can be natural use of value judgment at definition of distances between parts of considered territory or journey time, on the basis of experience of the expert with use of linguistic variables.

We call a variable as linguistic if its values are words or offers natural or an artificial language. So the concept "distance" is a linguistic variable if it accepts linguistic, instead of numerical values. For example, values of type far, not far, very far, quite normal, etc. (Zadeh 1975a; Zadeh 1975b; Zadeh 1975c).

We call a linguistic variable as syntactically-dependent if a procedure of formation of new values does not

depend on set of base values (Malishev et al. 1991). However there is a class of linguistic variables at which procedure of formation of new values depends on set of base values, and from a range of definition. So the linguistic variable «journey time» can have any values of type «about 50 minutes», «approximately 20 - 25 minutes». The variables can be considered as fuzzy number \tilde{a} and fuzzy interval $[\tilde{a}, \tilde{b}]$ with triangular and trapezoidal membership functions accordingly. We call such linguistic variables as syntactic-independent (Malishev et al. 1991). Semantic of values of syntactically-independent linguistic variable, i.e. membership functions are defined by definitional domain.

In this work, the approach to service centers allocation in a case when journey time or distance between parts of considered territory are sets of a syntactic-independent linguistic variable with trapezoidal membership functions is considered.

OPTIMUM ALLOCATION OF CENTERS BY FUZZY INTERVAL GRAPHS

We consider some territory which is divided into n areas. There are k service centres, which may be placed into these areas ($k < n$). It is supposed that service centre may be placed into some stationary place of each area. It is necessary for the given number of the service centers to define the places of their best allocation. In other words, it is necessary to define the places of k service centers into n areas such that the service of all territory was carried out on its minimum possible time or distance at least to one service center.

Let's consider, that the information received from GIS is presented in the form of the fuzzy interval graph $\tilde{G} = (X, \tilde{U})$. A set $X = \{x_i\}$, $i \in I = \{1, 2, \dots, n\}$ is the set of vertices. The vertices represent areas of some territory. A set $\tilde{U} = \{< \tilde{l}_{ij} / (x_i, x_j) >\}$ is the set of the fuzzy directed edges. A value $\tilde{l}_{ij} = [\tilde{a}_{ij}, \tilde{b}_{ij}]$ is fuzzy interval «approximately $[a_{ij}, b_{ij}]$ ». It is a meaning of syntactic-independent linguistic variable «time of journey from vertex x_i to vertex x_j ». Here $a_{ij}, b_{ij} \in R^1$, and $a_{ij} \leq b_{ij}$.

Let's believe, that the interval $\tilde{l}_{ii} = [0, 0]$, $\forall i \in \{1, 2, \dots, n\}$. In work (Dziouba and Rozenberg 2001) the method of a finding of optimum allocations of the service centers has been considered to crisp interval graphs. By means of this method for the set interval graph and the set number of the service centers k their optimum allocation is defined. However, it would be useful to obtain the given characteristics not for one centers of service in advance set number k , but for any number $k \in C[1, n]$. It would allow solve more objectively a problem of choice of service centers number.

For the decision of this problem we will consider the concept of interval base of the fuzzy interval graph. The given concept is an expansion of base for crisp and

fuzzy graphs (Bershtein and Bozhenuk 2008; Bershtein et al 2013).

Let $\tilde{l}_1 = [\tilde{a}_1, \tilde{b}_1]$ and $\tilde{l}_2 = [\tilde{a}_2, \tilde{b}_2]$ are any fuzzy intervals. We call the intervals \tilde{l}_1 and \tilde{l}_2 as incommensurable intervals if $a_1 > a_2$ and $b_1 < b_2$ are carried out. Otherwise, we can set naturally relations $>$, $<$, \leq , and \geq between fuzzy intervals. The sum of fuzzy intervals \tilde{l}_1 and \tilde{l}_2 we call an interval $\tilde{l} = [\tilde{a}, \tilde{b}]$, in which borders $a = a_1 + a_2$ and $b = b_1 + b_2$ (Hansen 1992).

We consider two operations $INCMIN(L)$ and $INCMAX(L)$ on set of intervals $L = \{l_i\}$. These operations are an estimation of subsets of the least and the greatest intervals from set of intervals L .

Example 1. Let set of intervals $L = \{[10, 15], [12, 14], [12, 17], [15, 18]\}$, then $INCMIN(L) = \{[10, 15], [12, 14]\}$ and $INCMAX(L) = \{[15, 18]\}$.

Let x and y are any vertices of fuzzy interval graph $\tilde{G} = (X, \tilde{U})$. We will define through \tilde{L}_{xy} a set of fuzzy intervals by means of which the vertex y is achievable from the vertex x . Then for each pair of vertices (x, y) we can put set of fuzzy intervals $INCMIN(\tilde{L}_{xy})$ in conformity.

We call a subset $B \subseteq X$ with family of intervals $\tilde{\Lambda}$ as interval base of fuzzy graph \tilde{G} , which is defined by expression:

$$\tilde{\Lambda} = INCMIN_{\forall y \in X \setminus B} \{ INCMAX_{\forall x \in B} \{ \tilde{L}_{xy} \} \},$$

and thus subset B is minimum in the sense that:

$(\forall B' \subset B) (\exists \tilde{l} \in \tilde{\Lambda}) (\exists \tilde{l}' \in \tilde{\Lambda}' \mid \tilde{l}' > \tilde{l})$, where family $\tilde{\Lambda}'$ is defined as:

$$\tilde{\Lambda}' = INCMIN_{\forall y \in X \setminus B'} \{ INCMAX_{\forall x \in B'} \{ \tilde{L}_{xy} \} \}.$$

Among all interval bases consisting of 1 vertex, we select such bases in which fuzzy intervals are the least or incommensurable. We designate them as $\tilde{\Lambda}_1$. Among all fuzzy interval bases consisting of 2 vertices we select such in which fuzzy intervals also are the least or incommensurable among themselves and we will designate them as $\tilde{\Lambda}_2$, and etc.

We call set $\tilde{B} = \{< \tilde{\Lambda}_1 / 1 >, < \tilde{\Lambda}_2 / 2 >, \dots, < \tilde{\Lambda}_n / n >\}$ fuzzy set of interval bases of graph \tilde{G} .

The fuzzy set of interval bases defines the set of the least or incommensurable fuzzy intervals (distance or time) from any area to some centre serving whole territory by k service centers ($k \in \{1, 2, \dots, n\}$).

Example 2. Consider the fuzzy interval graph presented in Figure 1:

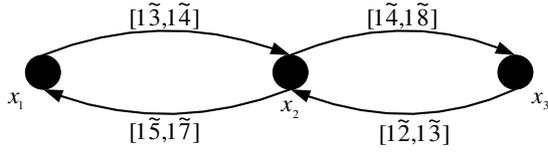


Figure 1: Fuzzy interval graph ($n=3$)

For the interval bases, consisting of 1, 2, and 3 vertices we will find the intervals families:

$$\begin{aligned} \tilde{\Lambda}_{11} &= \{[27, \tilde{3}2]\}, \text{ if } B_{11}=\{x_1\}; \\ \tilde{\Lambda}_{12} &= \{[15, \tilde{1}7], [14, \tilde{1}8]\}, \text{ if } B_{12}=\{x_2\}; \\ \tilde{\Lambda}_{13} &= \{[27, \tilde{3}0]\}, \text{ if } B_{13}=\{x_3\}; \\ \tilde{\Lambda}_{21} &= \{[12, \tilde{1}3]\}, \text{ if } B_{21}=\{x_1, x_3\}; \\ \tilde{\Lambda}_{22} &= \{[14, \tilde{1}8]\}, \text{ if } B_{22}=\{x_1, x_2\}; \\ \tilde{\Lambda}_{23} &= \{[15, \tilde{1}7]\}, \text{ if } B_{23}=\{x_2, x_3\}; \\ \tilde{\Lambda}_{31} &= \{[0,0]\}, \text{ if } B_{31}=\{x_1, x_2, x_3\}. \end{aligned}$$

Hence, the fuzzy set of interval bases is defined as:

$$\tilde{B} = \{ \langle [15, \tilde{1}7], [14, \tilde{1}8] \rangle / 1, \langle [12, \tilde{1}3] \rangle / 2, \langle [0,0] \rangle / 3 \}.$$

We will consider the method of finding a fuzzy set of interval bases. The given method is an analogue Maghout's method for definition of all minimal fuzzy dominating vertex sets (Bershtein and Bozhenuk 2001), and to Maghout's method for the definition of fuzzy vitality sets for fuzzy no interval graphs (Bozhenuk and Rozenberg 2012).

Let's consider some interval base B with a family of fuzzy intervals $\tilde{\Lambda}$. Let some interval \tilde{l} belongs to family $\tilde{\Lambda}$. Then for an arbitrary vertex $x_i \in X$, one of the following conditions must be true:

- $x_i \in B$;
- if $x_i \notin B$, then there is a vertex $x_j \in B$ such that fuzzy interval $\tilde{l}_{ji} = \tilde{l}(x_j, x_i)$ is no more \tilde{l} .

In other words, the following statement is true:

$$(\forall x_i \in X)(x_i \in B \vee (\exists x_j \in B | \tilde{l}_{ji} \leq \tilde{l})) \quad (1)$$

To each vertex $x_i \in X$ we assign Boolean variable p_i that takes the value 1, if $x_i \in B$ and 0 otherwise. Let's enter the predicate form $Q(\tilde{l}_{ji})$ that takes the value 1, if $\tilde{l}_{ji} \leq \tilde{l}$ and 0 otherwise. Using analogy between generality and existence quantifiers on the one hand, both operations conjunction and disjunction with another, we obtain a true logical proposition:

$$\Phi_B = \&_{i=1,n} (p_i \vee \vee_{j=1,n} (p_j \& Q(\tilde{l}_{ji}))) = 1. \quad (2)$$

Believing, that $Q(\tilde{l}_{ii}) = Q([0,0]) = 1$, an expression (2) may be rewrite as:

$$\Phi_B = \&_{i=1,n} (\vee_{j=1,n} (p_j \& Q(\tilde{l}_{ji}))) = 1. \quad (3)$$

We open the parentheses in the expression (3) and reduce the similar terms the following rules:

$$\left\{ \begin{aligned} &Q(\tilde{l}_1) \& Q(\tilde{l}_2) = Q(\tilde{l}_1), \text{ if } \tilde{l}_1 \geq \tilde{l}_2; \\ &p_1 \& p_2 \& Q(\tilde{l}_1) \& Q(\tilde{l}_2) \vee p_1 \& p_2 \& Q(\tilde{l}_3) = \\ &= p_1 \& p_2 \& Q(\tilde{l}_1) \& Q(\tilde{l}_2) \text{ if } \tilde{l}_1 < \tilde{l}_3 \& \tilde{l}_2 < \tilde{l}_3; \\ & \quad p_1 \& p_2 \& Q(\tilde{l}_1) \vee p_1 \& Q(\tilde{l}_2) = \\ & = p_1 \& Q(\tilde{l}_2), \text{ if } \tilde{l}_1 \geq \tilde{l}_2. \end{aligned} \right. \quad (4)$$

As a result expression (3) will become:

$$\Phi_B = \vee_{i=1,t} (p_{i1} \& p_{2i} \& \dots \& p_{ki} \& \& Q(\tilde{l}_{i1}) \& Q(\tilde{l}_{2i}) \& \dots \& Q(\tilde{l}_{ri})) = 1 \quad (5)$$

Property. If in expression (5) further simplification on the basis of rules (4) is impossible, then everyone disjunctive member i defines base with family of incommensurable intervals which are set by present predicates.

The following method of foundation of a fuzzy set of interval bases may be propose on the base of Property:

- we write proposition (3) for given fuzzy interval graph \tilde{G} ;
- we simplify proposition (3) by proposition (4) and present it as proposition (5);
- we define fuzzy set of interval bases, which correspond to the disjunctive members of proposition (5).

Example 3. Let's consider the given method on an example of the fuzzy interval graph presented in Figure 2:

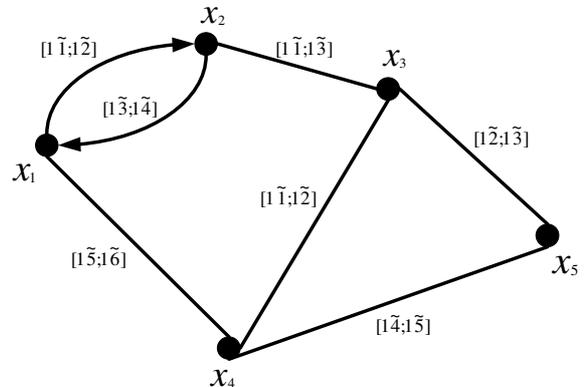


Figure 2: Fuzzy interval graph ($n=5$)

The adjacency matrix for this graph has the following form:

$$R_X = \begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{pmatrix} \infty & [1\tilde{1},1\tilde{2}] & \infty & [1\tilde{5},1\tilde{6}] & \infty \\ [1\tilde{3},1\tilde{4}] & \infty & [1\tilde{1},1\tilde{3}] & \infty & \infty \\ \infty & [1\tilde{1},1\tilde{3}] & \infty & [1\tilde{1},1\tilde{2}] & [1\tilde{2},1\tilde{3}] \\ [1\tilde{5},1\tilde{6}] & \infty & [1\tilde{1},1\tilde{2}] & \infty & [1\tilde{4},1\tilde{5}] \\ \infty & \infty & [1\tilde{2},1\tilde{3}] & [1\tilde{4},1\tilde{5}] & \infty \end{pmatrix} \end{matrix}.$$

For a finding of a reachability matrix of the graph, we will define operation of adjacency matrix exponentiation as:

- zero degree –

$$R_X^0 = \begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{pmatrix} [0,0] & \infty & \infty & \infty & \infty \\ \infty & [0,0] & \infty & \infty & \infty \\ \infty & \infty & [0,0] & \infty & \infty \\ \infty & \infty & \infty & [0,0] & \infty \\ \infty & \infty & \infty & \infty & [0,0] \end{pmatrix} \end{matrix}.$$

- second degree - $R_X^2 = R_X \times R_X = \{ \tilde{l}_{ik}^{(2)} \}$, where matrix elements are defined by formula:

$$\{ \tilde{l}_{ik}^{(2)} \} = \underset{j=1,n}{\text{INCMIN}} \{ \tilde{l}_{ij}^{(1)} + \tilde{l}_{jk}^{(1)} \};$$

- degree t - $R_X^t = R_X^{t-1} \times R_X$.

We define matrices $R_X^0, R_X^1, R_X^2, R_X^3, R_X^4$. Then we find their crossing. As a result we receive a reachability matrix

$$N_X = R_X^0 \cap R_X^1 \cap R_X^2 \cap R_X^3 \cap R_X^4 =$$

$$= \begin{matrix} & x_1 & x_2 & x_3 & x_4 & x_5 \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{pmatrix} \{ [0,0] \} & \{ [1\tilde{1},1\tilde{2}] \} & \{ [2\tilde{2},2\tilde{5}] \} & \{ [1\tilde{5},1\tilde{6}] \} & \{ [2\tilde{9},3\tilde{1}] \} \\ \{ [1\tilde{3},1\tilde{4}] \} & \{ [0,0] \} & \{ [1\tilde{1},1\tilde{3}] \} & \{ [2\tilde{2},2\tilde{5}] \} & \{ [2\tilde{3},2\tilde{6}] \} \\ \{ [2\tilde{4},2\tilde{7}] \} & \{ [1\tilde{1},1\tilde{3}] \} & \{ [0,0] \} & \{ [1\tilde{1},1\tilde{2}] \} & \{ [1\tilde{2},1\tilde{3}] \} \\ \{ [1\tilde{5},1\tilde{6}] \} & \{ [2\tilde{2},2\tilde{5}] \} & \{ [1\tilde{1},1\tilde{2}] \} & \{ [0,0] \} & \{ [1\tilde{5},1\tilde{5}] \} \\ \{ [2\tilde{9},3\tilde{1}] \} & \{ [2\tilde{3},2\tilde{6}] \} & \{ [1\tilde{2},1\tilde{3}] \} & \{ [1\tilde{4},1\tilde{5}] \} & \{ [0,0] \} \end{pmatrix} \end{matrix}.$$

Let's make an expression (2) on the received reachability matrix:

$$\begin{aligned} \Phi_B = & \{ Q(0;0)p_1 \vee Q(13;14)p_2 \vee Q(24;27)p_3 \vee \\ & \vee Q(15;16)p_4 \vee Q(29;31)p_5 \} \& \\ & \& \{ Q(11;12)p_1 \vee Q(0;0)p_2 \vee Q(11;13)p_3 \vee \\ & \vee Q(22;25)p_4 \vee Q(23;26)p_5 \} \& \\ & \& \{ Q(22;25)p_1 \vee Q(11;13)p_2 \vee Q(0;0)p_3 \vee \\ & \vee Q(11;12)p_4 \vee Q(12;13)p_5 \} \& \\ & \& \{ Q(15;16)p_1 \vee Q(22;25)p_2 \vee Q(11;12)p_3 \vee \\ & \vee Q(0;0)p_4 \vee Q(14;15)p_5 \} \& \\ & \& \{ Q(29;31)p_1 \vee Q(23;26)p_2 \vee Q(12;13)p_3 \vee \\ & \vee Q(14;15)p_4 \vee Q(0;0)p_5 \}. \end{aligned}$$

We open parentheses, reduce similar terms by the rules (4). We obtain finally:

$$\begin{aligned} \Phi_B = & Q(29;31)p_1 \vee Q(23;26)p_2 \vee Q(22;25)p_4 \vee \\ & Q(29;31)p_5 \vee Q(23;26)p_1p_2 \vee \underline{Q(22;25)p_1p_4} \vee \\ & Q(22;25)p_1p_5 \vee \underline{Q(12;13)p_1p_3} \vee Q(14;15)p_1p_4 \vee \\ & Q(14;15)p_1p_5 \vee Q(13;14)p_2p_3 \vee Q(14;15)p_2p_4 \vee \\ & Q(15;16)p_3p_4 \vee Q(11;12)p_1p_3p_5 \vee \underline{Q(11;12)p_1p_4p_5} \vee \\ & Q(15;16)p_1p_2p_3 \vee Q(15;16)p_1p_2p_4 \vee \underline{Q(0;0)p_1p_2p_3p_4p_5}. \end{aligned}$$

From this equation follows, that the fuzzy set of interval bases is:

$$\tilde{B} = \{ \langle [2\tilde{2};2\tilde{5}]/1 \rangle; \langle [1\tilde{2};1\tilde{3}]/2 \rangle; \langle [1\tilde{1};1\tilde{2}]/3 \rangle; \langle [0;0]/5 \rangle \}.$$

The fuzzy set of interval bases defines following optimum allocation of the centers: If we have 5 centers we place them in each vertex. In this case any expenses for achievement of other areas do not required (time equally 0). If we have 3 (or 4) centers they should be placed in the vertices 1, 4, and 5. In this case the least time is placed into fuzzy interval $[1\tilde{1};1\tilde{2}]$. If we have 2 centers they should be placed in the vertices 1 and 3. In this case the least time is placed into fuzzy interval $[1\tilde{2};1\tilde{3}]$. And at last if we have only 1 center it should be placed in the vertex 4. In this case the least time is placed in fuzzy interval $[2\tilde{2};2\tilde{5}]$.

MEMBERSHIP FUNCTION CALCULATION OF FUZZY INTERVALS

Let's define membership functions of the received fuzzy intervals. For this purpose we will take advantage of a method offered in work [9]. Let fuzzy time «near t' » is between the next base values «near t_1 » and «near t_2 » ($t_1 \leq t' \leq t_2$) which membership functions $\mu_{t_1}(t)$ and $\mu_{t_2}(t)$ look like a triangle. Then borders of membership function $\mu_{t'}(t)$ of fuzzy time «near t' » will set a linear combination of borders of the left and right base values:

$$l^L = \frac{(t_2 - t)}{(t_2 - t_1)} \times l_1^L + \left(1 - \frac{(t_2 - t)}{(t_2 - t_1)}\right) \times l_2^L, \text{ and}$$

$$l^R = \frac{(t_2 - t)}{(t_2 - t_1)} \times l_1^R + \left(1 - \frac{(t_2 - t)}{(t_2 - t_1)}\right) \times l_2^R.$$

It is schematically shown on Figure 3:

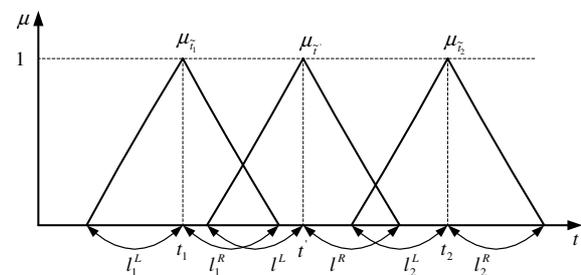


Figure 3: Definition of membership function $\mu_{t'}(t)$

Example 4. Let membership functions of base values of a syntactically-independent linguistic variable «time in ways» are presented in Figure 4 by fuzzy times «near 9», «near 12», «near 18», «near 23», «near 25» and «near 30».

So, we calculate $l^L = 2, l^R = 2$ for number $1\tilde{1}$; similarly for number $1\tilde{3}$ it is found $l^L = 2,2; l^R = 2,2$; and for number $2\tilde{2}$ it is found $l^L = 3, l^R = 3$.

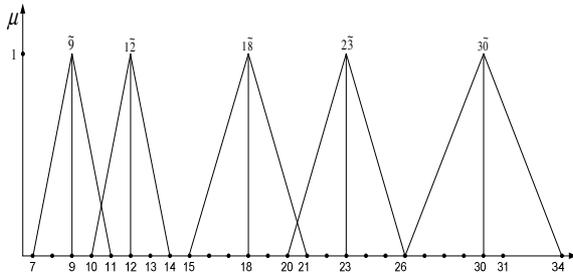


Figure 4: Membership functions of base values of a syntactically-independent linguistic variable «time in ways»

The given approach allows define parameters of membership function $\mu_{[\tilde{x}_l, \tilde{x}_r]}(x)$ of any fuzzy interval $[\tilde{x}_l, \tilde{x}_r]$ in the trapezium form with the calculated borders:

$$\mu_{[\tilde{x}_l, \tilde{x}_r]}(x) = \begin{cases} \mu_{\tilde{x}_l}(x), & \text{if } x \leq x_l; \\ 1, & \text{if } x_l \leq x \leq x_r; \\ \mu_{\tilde{x}_r}(x), & \text{if } x \geq x_r. \end{cases}$$

Differently, the method allows to calculate inclinations of the left and right arms of trapezes which correspond to linguistic concepts «about an interval 22-25», «about an interval 12-13» and «about an interval 11-12», as is shown in Figure 5:

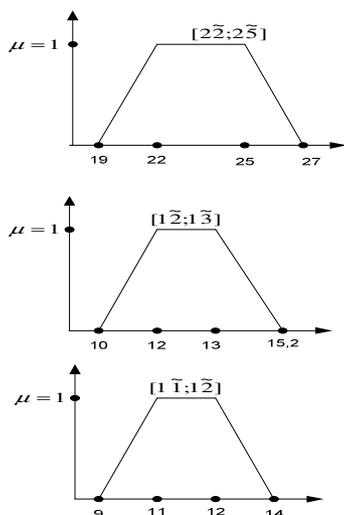


Figure 5: Membership functions of concepts «about an interval 22-25», «about an interval 12-13» and «about an interval 11-12»

Having similar idea of the finding results, it is possible to speak about interval time between objects at those or other values of membership function.

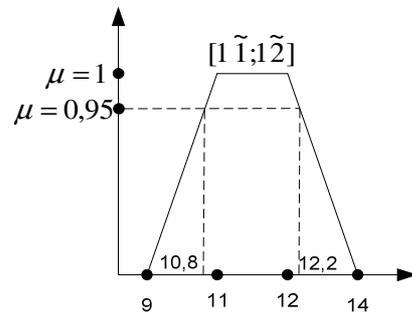


Figure 6: The least time for membership function not less than 0.95

So, for membership function not less than 0.95, the least time value will be within interval $[10.8, 12.2]$ in the case of existing of two service centers, as is shown in Figure 6.

CONCLUSION

The task of defining of optimal allocation of centres as the task of definition fuzzy bases of fuzzy interval graphs was considered. The definition method of fuzzy bases is the generalization of Maghout's method for fuzzy no interval graphs. This method is effective for the graphs which have no homogeneous structure and no large dimensionality. In our future work we are going to investigate the task of the service centers accommodation, based on the fuzzy set of antibases rather than fuzzy set of bases. The problems of the centers location on the fuzzy temporal graphs, i.e. such graphs, whose edges membership functions change in discrete time, will also be studied. It is necessary to notice, that the considered method allows define the best places of allocations of the service centers in case of their placing only in vertices of the graph (instead of on edges with generation of new vertices). Besides, certain interest is caused by a problem of convolution of incommensurable intervals that would allow simplify process of calculations.

Acknowledgements. This work was partially supported by the Russian Foundation for Basic Research, projects N 13-07-13103 ofi_m_RzhD, and N 14-01-00032.

REFERENCES

- Bershtein, L. and A. Bozhenuk. 2001. "Maghout method for determination of fuzzy independent, dominating vertex sets and fuzzy graph kernels", *J. General Systems*, vol. 30, pp.45-52.
- Bershtein, L. and A. Bozhenuk. 2008. "Fuzzy graphs and fuzzy hypergraphs". In J. Dopico, J. de la Calle, and A. Sierra, editors, *Encyclopedia of Artificial Intelligence Information SCI*, pages 704-709, Hershey, New York.

- Bershtein, L.; A. Bozhenyuk; and I. Rozenberg. 2013. "Optimum allocation of centers in transportation networks by means of fuzzy graph bases". In *Proceedings of the 8th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13). Advances in Intelligent Systems Research*. Editors: Gabriella Pasi, Javier Montero, Davide Ciucci, 11 – 13 September 2013. Atlantis Press, pp.230-235.
- Bozhenyuk, A. and I. Rozenberg. 2012. "Allocation of service centers in the GIS with the largest vitality degree", In S. Greco et al., editors, IPMU 2012, Part II, *Communications in Computer and Information Science*, CCIS 298, Springer-Verlag, Berlin Heidelberg, pp.98-106.
- Christofides, N. 1976. *Graph Theory. An Algorithmic Approach*. Academic press, London.
- Clarke, K. and C. Englewood. 1995. *Analytical and Computer Cartography*. Prentice Hall, USA.
- Dziouba, T. and I. Rozenberg. 2001. "The decision of service centres location problem in fuzzy conditions", In: Bernd, Reusch (ed.): *Lecture Notes in Computer Science*, Vol. 2206, *Computational Intelligence: Theory and Applications. Proceedings of International Conference 7th Fuzzy Days*, Springer Verlag, pp.11-17.
- Goodchild, M. 1989. "Modelling Error in Objects and Fields". In: Goodchild, M.F., Gopal, S. (eds.): *Accuracy of Spatial Databases*. Taylor & Francis, Inc., New York, USA.
- Hansen, E. 1992. *Global Optimization Using Interval Analysis*. Dekker, New York, USA.
- Kaufmann, A. 1977 *Introduction a la Theorie des Sous-ensembles Flous*, Masson, Paris, France.
- Longley, P., M. Goodchild, D. Maguire, D. Rhind. 2001. *Geographic Information Systems and Science*. John Wiley & Sons, Inc., New York, USA.
- Malczewski, J. 1999. *GIS and Multicriteria Decision Analysis*. John Wiley & Sons Inc., New York, USA.
- Malishev, N.G.; L.S. Bershtein; and A.V. Bozhenyuk. 1991. *Fuzzy Models for Expert Systems in Computer Aided Design*. Energoatomizdat, Russia.
- Zadeh, L. 1975a. "The concept of a linguistic variable and its application to approximate reasoning", Part I, *Information Sciences*, no. 8, pp.199-249.
- Zadeh, L. 1975b. "The concept of a linguistic variable and its application to approximate reasoning", Part II, *Information Sciences*, no. 8, pp.301-357.
- Zadeh, L. 1975c. "The concept of a linguistic variable and its application to approximate reasoning", Part III, *Information Sciences*, no. 9, pp.43-80.
- Zhang, J. and M. Goodchild. 2002. *Uncertainty in Geographical Information*. Taylor & Francis, Inc., New York, USA.

AUTHOR BIOGRAPHIES



LEONID S. BERSHTEIN was born in Kiev, Ukraine and went to the Radio Engineering Institute of Taganrog, where he studied and obtained his degree in 1983. He is scientific Advisers of Scientific and Technical Center "Information Technologies", Southern Federal University. His research interests include fuzzy logic, fuzzy graphs and hypergraphs. He has more than 400 publications in the fields of Computer Sciences. Contact information: 44, Nekrasovsky Street, Taganrog, 347900,

Russia, phone: +78634371695, e-mail: lsbershtyayn@sfnedu.ru.



ALEXANDER V. BOZHENYUK was born in Taganrog, Russia. Professor of Scientific and Technical Center "Information Technologies", South Federal University, Russia. He holds a degree of Doctor of Technical Sciences in Theoretical Foundations of Informatics in 2001. His research interests include fuzzy models, fuzzy decision making. He has more than 180 publications in the fields of Computer Sciences. Contact information: 44, Nekrasovsky Street, Taganrog, 347900, Russia, phone: +78634371743, e-mail: avb002@yandex.ru.



STANISLAV L. BELYAKOV was born in Ukraine. He graduated Leningrad Electrotechnical Institute in 1982. Professor of Scientific and Technical Center "Information Technologies", South Federal University, Russia. He holds a degree of Doctor of Technical Sciences in Theoretical Foundations of Informatics in 2003. His research interests include geographic information systems, modeling intelligence. Contact information: 44, Nekrasovsky Street, Taganrog, 347900, Russia, phone: +78634371743, e-mail: beliacov@yandex.ru.



IGOR N. ROZENBERG was born in Taganrog, Russia and went to the Radio Engineering Institute. He holds a degree of Doctor of Technical Sciences in Geographic Information Systems in 2007. He is the First Deputy General Director of Public Corporation "Research and Development Institute of Railway Engineers". Contact information: 27/1, Nizhegorodskaya Street, Moscow, 109029, Russia, phone: +74959677701, e-mail: I.rozenberg@gismps.ru.

APPROACHES TO RUN SIMULATIONS OF BUSINESS PROCESSES IN A GRID COMPUTING NETWORK

Christian Müller
 Department of Management and Business Computing
 Technical University of Applied Sciences Wildau
 Hochschulring 1
 D-15745 Wildau, Germany
 christian.mueller@th-wildau.de

KEYWORDS

Event driven simulation, business processes, grid computing

ABSTRACT

Two approaches to run simulations of business processes in a grid computing network are compared. Based on special properties of simulation models, the approach that use a grid framework is more stable than the web service approach.

INTRODUCTION

Business processes of modern companies are characterized by a huge complexity which is caused for example by quickly changing markets, short product life cycles or dynamic interactions between particular subsystems of a company. Business process management is intended to implement efficient and customer-oriented processes whereby the simulation of business processes can be used to evaluate the quality of processes and to identify areas of improvements. To analyze these models we must run many of experiments, which needs a lot of time on a single computer. In this paper we will discuss two approaches to parallelize the running of the experiments in a grid network. For modeling of business processes as an event driven simulation model we use the Epc-Simulator as simulation system (Müller 2012), (Müller 2014).

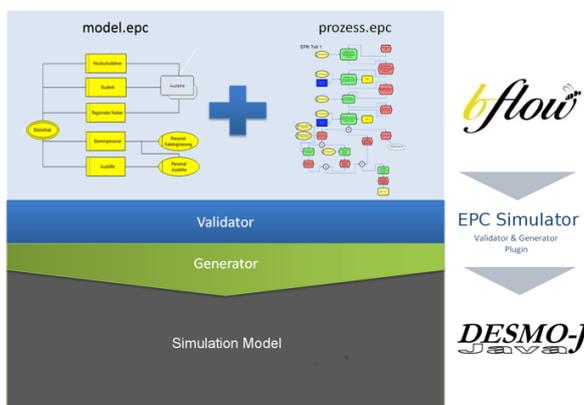


Figure 1: Epc-Simulator concept image

The Epc-Simulator is a plugin of the EPC modeling

toolbox bflow* (Kern et al. 2010), (Bflow 2014). As shown in Figure 1, the first step to create a simulation model is the specification of a model document and several process documents in bflow*. The model document contains information about the model infrastructure (simulation time, available resources, inter-arrival times of entities, etc.). The process documents contain the process descriptions in EPC notation. Based on these documents, the Epc-Simulator can generate a simulation model. This is a Java Application that uses the DESMO-J Framework, whereby DESMO-J provides the basic functionality of a simulation. (Page and Kreutzer 2005), (DesmoJ 2014)

Each business process described with Epc-Simulator contains a model and some process documents. The idea of modeling with Epc-Simulator will be introduced with the following supermarket example.

A supermarket is entered by customers. The inter-arrival time between two customers that enter the shop is given by a probability distribution. The process that each customer runs is described in a process document named customer (Figure 3). In the supermarket work one or more entities named cashier. This is described in the model document in Figure 2.

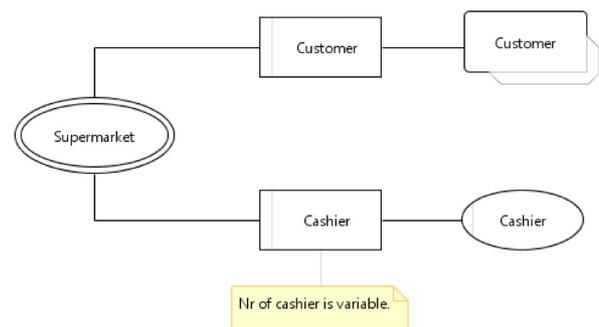


Figure 2: Model document

Every model is parametrized by some standard parameters, like a seed value for the random generator and some optional model-dependent parameters, like the number of available cashiers. On the other hand the simulator computes some performance indices from each simulation run, e.g. the utilization rate of the cashiers.

The Epc-Simulator produces an executable jar archive with the generated simulation model and a parameter file. Before the simulation runs, this file contains the input data and after the run the file is extended by the output data.

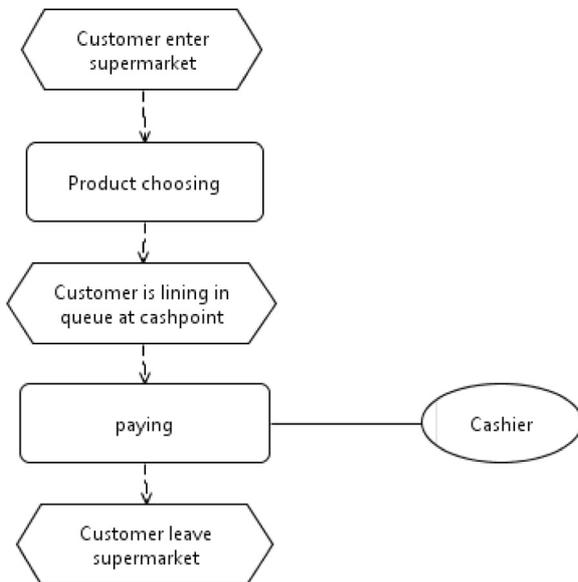


Figure 3: Process document

Before running a row of experiments, the parameter file for each experiment must be customized with an individual parameter file and specialty with its own seed value.

STRUCTURES OF DISTRIBUTED COMPUTING NETWORKS

Distributed computing networks can be organized as Peer-to-Peer, as Cluster or Grid networks. In a Peer-to-Peer network (Figure 4), all nodes in the network are equal and each node can cooperate with an other. This can produce a huge organization overhead. For this see (Bengel 2004), (Bengel et. al. 2008), (Dunkel 2008, (Schill 2012).

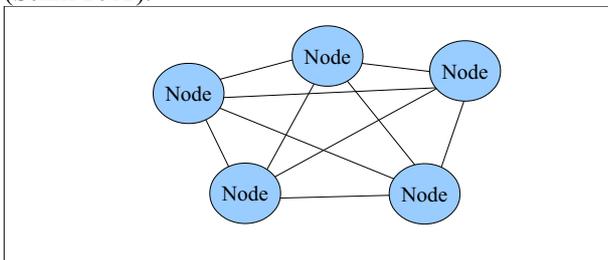


Figure 4: Peer to Peer Network

In a Cluster (Figure 5) we have a client and he can distribute his tasks to nodes that are associated with him exclusively. A cluster works inside of an organization and has no access to external nodes. On the other hand in a grid network a client works with nodes that can be located everywhere. For this, the grid network uses an open protocol stack.

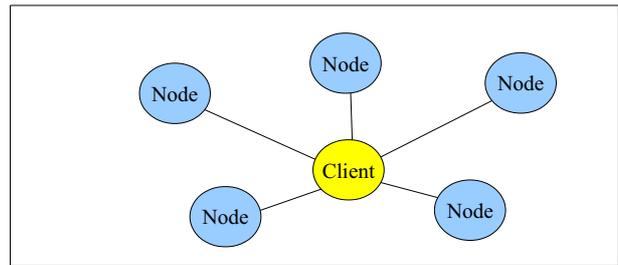


Figure 5: Cluster Network

In this paper we compare two approaches to distribute simulation tasks in a cluster. For this we use techniques that are working on a high abstraction level. The first is a web service approach and the second uses a grid framework. Both approaches are tested only in a cluster, but the principles there are working in a grid too. Approaches that are working on a low abstraction level, like RMI, CORBA or RPC are not examined.

SIMULATION TASKS AS DISTRIBUTED SOFTWARE OBJECTS

A simulation job contains a common jar archive with the simulation model and a parameter file for each simulation task. The client distribute the tasks of a job on a set of nodes. The following executing properties of a task depend on its parameter file and are in general not predictable:

- The *elapsed time of the simulation* depends on the length of the simulation period and the event frequency of the model. The frequency increases with the model complexity.
- The *memory (RAM) requirement of the simulation* model depends on the individual model properties. When e.g. in the example above the inter-arrival times of the customer are shorter than the service times of the cashier, then the queue in front of the sale station will increase. Consequently, there are more entities in the simulation and the simulation model needs more memory. In complex models it is not easy to extrapolate the memory requirement.
- The simulation model can produce a *data file to animate the simulation*. The length of this file depends on the length of the simulation period and the event frequency of the model. This file can be big. Hence the size of response data to send from node to client can be big.

In these points the properties of simulation models differ from common distributed software objects, like Mandelbrot sets. There are fast, short and their response is only a number. On the other hand, there are more distributed computations than simulation experiments in our case.

USING WEB SERVICES TO ORGANIZE THE GRID

In this approach, the simulation models are enveloped in a wrapper that includes web service functionality. We have developed a wrapper for the REST and one for the SOAP protocol. On each node runs a servlet engine and the models must be deployed on the servlet engine of this node. The client sends simulation tasks, together with their parameter files to the nodes. A manager of the node collects all simulation tasks and organizes the sequence of their execution. The client asks after a while about the status of his tasks. When a task is completed, it can download the results.

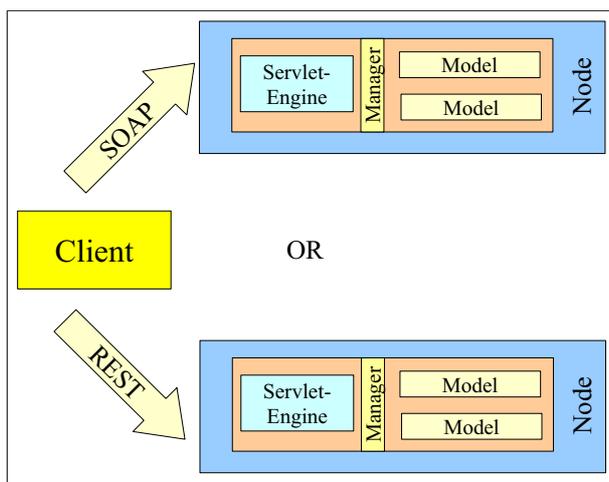


Figure 6: Using web services

The servlet engine, here we used a Glassfish, and their deployed applications, here the simulation models, are running in the same java virtual machine. Sometimes, when the simulation model is badly parametrized, the model runs in an out of memory exception. In consequence of that, the servlet engine runs after the next client access also in a exception and hence the node is no longer accessible from the client (Zimmermann et. al. 2012).

Its not unlikely that a simulation model breaks with a out of memory exception. We will illustrate this on the example above. When the model runs over a long time and more customers enter than leave the shop, then the queue in front of the cashier will go longer and longer. Since every customer entity needs a bit of memory, the model requires after a while more memory than accessible.

USING THE GRID FRAMEWORK-JPPF TO ORGANIZE THE GRID

As an alternative for the web service approach to distribute simulation tasks in a grid, we use JPPF as a grid framework. JPPF stands here as an example of a light weight grid framework(Jensen 2013), (JPPF 2014). JPPF is a open source framework and runs with the Apache 2 licence (Apache Licence 2004). An overview

and comparison of such frameworks is given in (Azadzadeh 2006)

A client application sends a job with a row of simulation tasks to a JPPF server. The server distributes the tasks to nodes. When the tasks on a node are completed the server collects the simulation results from the nodes and delivers them to the client application. On the server and on all nodes runs a small sized JPPF management software. When in a simulation task an exception occurs, the simulation is finished and is marked as not runnable. In our experiments the break of a simulation model has no side effect on the stability of the node management software.

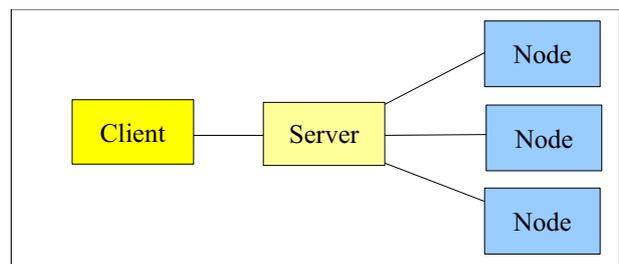


Figure 7: Using a grid framework

In experiments with 1000 simulation tasks and 5 nodes, all results arrived the client nearly at the same time. Hence, the result handling in the client can be a bottleneck. To avoid this, we can use some JPPF notification mechanism to inform the client when a single task is finished. With this information we can distribute the activities in the client to handle the results on the whole processing time of the distributed simulation.

CONCLUSION

In this paper we compared two approaches to distribute simulation tasks in a grid network. One approach uses a web service to organize the communication between client and computing nodes and the other uses the grid framework JPPF.

For the web service approach, on each computing node runs a servlet engine. This is a heavy-sized application. When the run of a simulation model breaks with an out of memory exception, the servlet engine may break by the same reason.

On the other hand, the distribution approach with the JPPF grid framework runs stably, also when the simulation model breaks with an out of memory exception. For the management of a node, there runs a light weight management software. The less memory requirements of the node management system may be a reason of their stability.

This investigation shows that the approach with the grid framework is more stable than the web service approach. In further investigations we will have a look

at other grid frameworks. Furthermore, more numerical experiments are required.

REFERENCES

- Asadzadeh, P; Buyya, R; Kei, C; Nayar, D; Venugopal, S:* Global Grids and Software Toolkits – A Study of Four Grid Middleware Technologies 2006 in High-Performance Computing: Paradigm and Infrastructure eds: Yang, L; Guo, M; John Wiley & Sons
- Apache License:* Apache License, Version 2.0, 2004
<http://www.apache.org/licenses/LICENSE-2.0.html>
- Bengel, G:* Grundkurs Verteilte Systeme 2004 Friedrich Vieweg & Sohn Verlag
- Bengel, G; Baum, C; Kunze, M; Stucky, K-U:* Masterkurs Parallele und Verteilte Systeme 2008 Vieweg +Teubener | GWV Fachverlage GmbH
- Bflow:* Bflow* Toolbox – Geschäftsprozessmodellierung als Open Source 2014 <http://www.bflow.org/>
- DesmoJ:* A Framework for Discrete-Event Modelling and Simulation 2014
<http://desmoj.sourceforge.net/home.html>
- Dunkel, J; Eberhard, A; Fischer, S; Kleiner, C; Koschel, A:* Systemarchitekturen für Verteilte Anwendungen 2008 Carl Hanser Verlag
- Heiko Kern, Stefan Kühne, Ralf Laue, Markus Nüttgens, Frank J Rump, Arian Storch:* bflow* Toolbox - an Open-Source Business Process Modelling Tool 2010, Proc. of BPM Demonstration Track 2010, Business Process Management Conference 2010 (BPM'10), Hoboken, USA
- Jensen, A:* Ein Ansatz zur Parallelisierung von Geschäftsprozessen mit dem Programm EPC Simulator an Beispiel des Grid-Frameworks JPPF

2013, TH Wildau Masterthesis

JPPF: JPPF Homepage 2014 <http://www.jppf.org/>

Müller, Chr : Generation of EPC Based Simulation Models 2012 In: Proceedings 26th European Conference on Modelling and Simulation 2012, Koblenz,
<http://dx.doi.org/10.7148/2012-0301-0305>

Müller, Chr. EpcSimulator Project Homepage 2014
<http://www.tfh-wildau.de/cmuller/EpcSimulator/>

Page, B; Kreutzer; W : The Java Simulation Handbook, Shaker 2005

Schill, A; Springer, T: Verteilte Systeme – Grundlagen und Basistechnologien 2012 Springer Verlag

Zimmermann, S; Sobottka, M; Szott, E: Fehlerbehebung in komplexen IT-Systemen am Beispiel von Simulation on Demand Webservices auf einer GlassFish-Serverumgebung 2012 Belegarbeit TH Wildau

AUTHORS BIOGRAPHIES



CHRISTIAN MÜLLER has studied mathematics at Free University Berlin. He obtained his PhD in 1989 about network flows with side constraints. From 1990 until 1992 he worked for Schering AG and from 1992 until 1994 for Berlin Public Transport (BVG) in the area of timetable and service schedule optimization. In 1994 he got his professorship for IT Services at Technical University of Applied Sciences Wildau, Germany. His research topics are conception of information systems plus mathematical optimization and simulation of business processes.

His email address is: christian.mueller@th-wildau.de and his web page is <http://www.th-wildau.de/cmuller/> .

Optimisation of Boids Swarm Model Based on Genetic Algorithm and Particle Swarm Optimisation Algorithm (Comparative Study)

Saleh Alaliyat
Faculty of Engineering and Natural
Sciences
Aalesund University College
N-6025, Aalesund, Norway
Email: saal@hials.no

Harald Yndestad
Faculty of Engineering and Natural
Sciences
Aalesund University College
N-6025, Aalesund, Norway
Email: hy@hials.no

Filippo Sanfilippo
Department of Maritime Technology
and Operations
Aalesund University College
N-6025, Aalesund, Norway
Email: fisa@hials.no

KEYWORDS

Flocking behaviour, genetic algorithms, and particle swarm optimisation.

ABSTRACT

In this paper, we present two optimisation methods for a generic boids swarm model which is derived from the original Reynolds' boids model to simulate the aggregate moving of a fish school. The aggregate motion is the result of the interaction of the relatively simple behaviours of the individual simulated boids¹. The aggregate moving vector is a linear combination of every simple behaviour rule vector. The moving vector coefficients should be identified and optimised to have a realistic flocking moving behaviour. We proposed two methods to optimise these coefficients, by using genetic algorithm (GA) and particle swarm optimisation algorithm (PSO). Both GA and PSO are population based heuristic search techniques which can be used to solve the optimisation problems. The experimental results show that optimisation of boids model by using PSO is faster and gives better convergence than using GA.

INTRODUCTION

Many animals in the nature move in groups: fish swim in schools, birds fly in flocks, sheep move in herds, insects move in swarm and ants distribute to find a food source and then all ants follow path to the food. Simulating the aggregate motion is an important issue in the areas of artificial life and computer animation and in a lot of their applications such as games and movies. Reynolds (1987) proposed a first computer model of group animal motion such as fish schools and bird flocks. He called his model as "boids model", where boids refer to the generic flocking simulated creatures. The aggregate moving of the simulated boids is the result of the interaction of the relatively simple behaviours of the individual simulated boid. An interesting look at boids can be taken from the perspective of artificial life where the holistic emergent phenomena is the result of interactions of independent entities (Anthony 2002). The boids model has three simple rules applied to the boids. The rules are: each boid move to avoid crowding with its

neighbours, match and coordinate its movements with its neighbours, and move to gather with the others. Other rules such as avoiding obstacles and goal seeking have been included in steering behaviour model (Reynolds, 1999) and in many simulations based on boids model later on. For instance, Delgado (2007) extended the basic boids model to include the effects of fear. Olfaction was used to transmit emotion between animals, through pheromones modelled as particles in a free expansion gas. Hartman and Benes (2006) added a complementary force to the alignment (steer towards the average heading of neighbours) that they call the change of leadership. This steer defines the chance of the boid to become a leader and try to escape.

After 1987, the boids model is often used in computer graphics to provide realistic life-like representations of the aggregate motion of groups. For instance, in the 1998 Valve Video Game Company has used boids model in Half-Life video game for the flying bird-like creatures (Valvesoftware.com 2014). The boids model represented an enormous step forward compared to the traditional techniques used in computer animation for motion pictures. The first animation created with the boids model was in the computer animation shot film called Stanley and Stella in: Breaking the Ice in 1987 (Ice' and Malone, 2014). After that, the boids model was used in a feature film introduction of Batman Returns in 1992 (Returns and Burton et al., 2014). Then the boids model has been used in many games and films and in many other interesting applications.

In the boids swarm model, each rule is represented by a vector. The vector by its two components (magnitude and direction) is adaptive to the environment. The boid moving vector is a linear combination of every behaviour rule vector. The setting of the moving vector coefficients becomes more difficult by increasing more behaviour rules. These coefficients should be determined and optimised to have a realistic moving behaviour. In this context, we use two optimisation algorithms to optimise the boids model by finding the best coefficients values and combination in order to minimise the objective function. Firstly, we propose a GA to optimise the coefficients in a generic boids model. Secondly we substitute the GA by PSO to optimise the coefficients in the same generic boids model and use PSO to find food sources. Then we do a comparison of these two models by focusing on the advantages and disadvantages of each algorithm.

¹ Boids are bird-like objects that were developed in the 1980s to model flocking behaviour.

THE BOIDS MODEL

In 1986, Reynolds has developed the boids model. His published paper about the boids model (Reynolds 1987) was cited so many times and extended in so many different ways. Many of the extensions present additional rules to the boids, some describe constrained solution, some tend to easy solutions usable in computer games, some extend the previous work in spite of computational complexity, etc.

Reynolds (1987) describes the flock behaviour² as a result of the motion and the interaction of boids. Each boid has three simple rules of steering behaviours that describe how an individual boid move based on the positions and velocities of its flock mates (social reaction).

- Separation (*figure 1(a)*): each boid keep a distance from other boids nearby to avoid collision and prevent crowding.
- Alignment (*figure 1(b)*): each boid match the direction and the speed of its neighbours. This rule causes boids to follow each other.
- Cohesion (*figure 1(c)*): each boid tends to move to the average position of its neighbours.

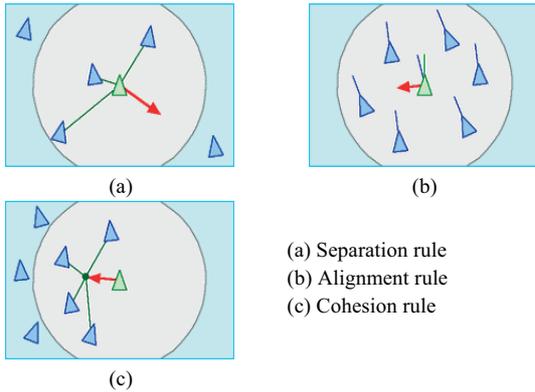


Figure 1: The boids social rules (Reynolds 1987).

Reynolds (1999) has extended the boids model to include more individual-based rules of the steering behaviours, to have more advanced individuals which are capable to finish specific task or adapt to complex environments. Some of these behaviours are:

- Obstacle avoidance (*figure 2(a)*): The obstacle avoidance behaviour allows the boids move in cluttered environment by dodging around obstacles.
- Leader following (*figure 2(b)*): this behaviour causes one or more boids to follow another moving boid selected as a leader.

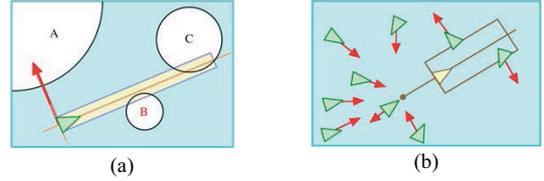


Figure 2: Steering behaviours rules (a) Obstacle avoidance (b) leader following(Reynolds 1999).

Based on Reynolds model, we have implemented a generic boids swarm model in Unity3D (Unity3d.com 2014), the program is written in MonoDevelop Unity-C# (Docs.unity3d.com 2014). The aim is to develop a generic model that can be used in simulating the aggregate motion for flocks of birds, schools of fish or herds of animals. The model has five rules:

- Cohesion: steer to move toward the average position of local flock mates (as in the original Reynolds' model). By applying cohesion rule keeps the boids together. This rule acts as the complement of the separation. If only cohesion rule is applied, all the boids in the flock will merge into one single position. Cohesion (Coh_i) of the boid (b_i) is calculated in two steps. First, the center ($\vec{F}c_i$) of the flock (f) that has this boid is calculated as in *equation 1*. Then the tendency of the boid to navigate toward the center of density of the flock is calculated as the cohesion displacement vector as in *equation 2*.

$$\vec{F}c_i = \sum_{v b_j \in f} \frac{\vec{p}_j}{N} \quad (1)$$

Where, p_j is the position of boid j and N is the total number of boids in f

$$\vec{Coh}_i = \vec{F}c_i - \vec{p}_i \quad (2)$$

- Alignment: steer to match the heading and the speed of its neighbours. This rule tries to make the boids mimic each other's course and speed. Boids tend to align with the velocity of their flock mates. The alignment (Ali_i) is calculated in two steps. First, the average velocity vector of the flock mates ($\vec{F}v_i$) is calculated by *equation 3*. Then Ali_i is calculated as the displacement vector in *equation 4*.

$$\vec{F}v_i = \sum_{v b_j \in f} \frac{\vec{v}_j}{N} \quad (3)$$

$$\vec{Ali}_i = \vec{F}v_i - \vec{v}_i \quad (4)$$

Where \vec{v}_i is the velocity vector of boid i
If this rule was not used, the boids would bounce around a lot and not form the beautiful flocking behaviour that can be seen in the nature.

- Separation: steer to avoid collection and overcrowding with other flock mates. There are many ways to implement this rule. An efficient solution to calculate the separation (Sep_i) is by applying *equation 5*. Vectors defined by the

² We mean by *Flock behaviour* in this paper as behaviour of flock, school, herd and swarm.

position of the boid b_i and each visible boid b_j are summed, then separation steer (\overrightarrow{Sep}_i) is calculated as the negative sum of these vectors.

$$\overrightarrow{Sep}_i = -\sum_{\forall b_j \in f} (\vec{p}_i - \vec{p}_j) \quad (5)$$

If only the separation rule is applied, the flock will dissipate.

- Leader following: steer to follow another moving boid selected as a leader (p_l). The leader following (\overrightarrow{Led}_i) is calculated by equation 6.

$$\overrightarrow{Led}_i = L * (\vec{p}_l - \vec{p}_i) \quad (6)$$

Where L is a leader strength factor. (Note: the moving vector (velocity) has limits, minimum and maximum).

- Random movement: this rule is added to have more realistic flock behaviour. This rule is depending on the random number generator inside the game engine (*Unity3D*). The random movement (\overrightarrow{Rand}_i) is calculated as in equation 7.

$$\overrightarrow{Rand}_i = -ffactor * \vec{r} \quad (7)$$

Where r is a unit sphere random vector and $ffactor$ is a flock random strength factor.

Then the moving vector (v_i) for boid (b_i) is calculated by combining all the steering behaviour vectors as in equation 8.

$$\vec{v}_i = w_1 \overrightarrow{Coh}_i + w_2 \overrightarrow{Ali}_i + w_3 \overrightarrow{Sep}_i + w_4 \overrightarrow{Led}_i + w_5 \overrightarrow{Rand}_i \quad (8)$$

Where w_i are the coefficients describing influences of each steering rule and used to balance the five rules.

We have used the *Unity3D* to implement the boids model to get the benefits of using a game engine. The first benefit is the amazing visualisation that we get in *Unity3D*. So we skip wasting time to model and program the boid's shape and its geometry. In *Unity3D*, it is easy to build a boid such as a bird, a fish or a sheep and attach some life-look animation to it, or import the 3D model of the boid from other programs and attach a built in animation to it or program the animation from the scratch. In this model we exploit the collision detection component in *Unity3D* game engine to avoid the obstacles. Each obstacle has a physic's collision component that doesn't let other objects to move through the collision bounds (obstacle's space). In another words, the obstacles will be excluded from the boids flocking space. We will show the results in the experimental results section.

GENETIC ALGORITHMS FOR OPTIMISATION OF BOIDS MODEL

In this section, we will give an overview of GA in general and some examples of its applications. Then we present our proposed model (GA for optimisation of boids model).

Genetic Algorithm:

Genetic algorithm (GA) is an optimisation and search technique based on the principles of genetic and natural selection (Haupt, 2003). A GA allows a population composed of many individuals to evolve under specified selection rules to a state that maximise the fitness (i.e., minimizes the cost function). Genetic algorithms (GAs) were invented by John Holland in 1960s and were developed by him and his students in 1960s and 1970s. Holland (1975) presented the GA as an abstraction of biological evolution and gave a theoretical framework for adaptation under GA.

GA belong to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution such as selection (reproduction), crossover (recombination) and mutation. The evolution process starts from a population of individuals generated randomly within the search space and continues for generations. In each generation, fitness of every individual is evaluated, and multiple individuals are randomly selected from the current population based on their fitness and modified by recombination and mutation operation to form a new population. Then this new population will be used for the next generation of the evolution. In general, the search process ends when either a maximum number of generations have been produced or a fitness level has been reached for the population. The flowchart of GA is shown in (figure 3).

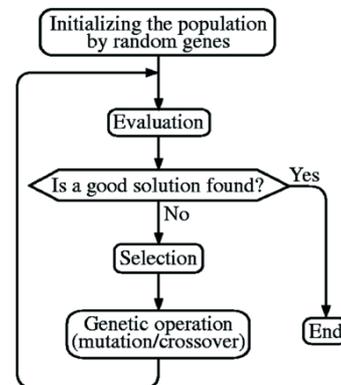


Figure 3: Flowchart of GA

In a GA, it's necessary to be able to evaluate how good a potential solution is relative to other potential solutions. The fitness function is responsible for performing this evaluation and returning a fitness value (positive integer number) that reflects how optimal the solution is. The fitness function is associated with the objective function of the problem. The fitness value of the individual is used to determine the probability with which the individual is selected into the new population. A common metaphor for the selection process is a roulette wheel selection (Fogel, 2000).

Traditional methods of search and optimization are too slow in finding a solution in a very complex search space. GA is a robust search method requiring little information to search effectively in a large or poorly understood search space. In particular a genetic search

progress through a population of points in contrast to the single point of focus of most search algorithms. Moreover, it is useful in the very tricky area of nonlinear problems. GAs have been used to solve optimisation problems in different fields such as automotive design, engineering design, robotics, optimised routing, games, etc. Chen (2006) has applied GAs to optimise the behaviour of a school of fish.

GA for optimisation of moving vector in the boids model:

The moving vector (v_i) in equation 8 for each boid (b_i) is a combination of all the five steering behaviour vectors. And the movements are balanced by the w_i weight coefficients, so these coefficients should be optimized to have realistic and life-look behaviour. We have removed the random steering behaviour in the moving vector to exclude the random movements for the boids. The new moving vector is:

$$\vec{v}_i = w_1 \vec{Coh}_i + w_2 \vec{Al}_i + w_3 \vec{Sep}_i + w_4 \vec{Led}_i \quad (9)$$

We have used GA to optimise these coefficients and getting the benefit from using GA for parameters optimisation and finding a global optimum solution. The goal is to find the optimal solutions in terms of the variables (coefficients). Thus we should define mathematically what is the optimal solution. We begin the GA by defining the chromosome. The chromosome is an array of the coefficients values that will be optimised. In this case the chromosome has four variables and is written as a four-element row vector.

$$chromosome = [w_1, w_2, w_3, w_4] \quad (10)$$

Then we should formulate the cost function that gives a cost for each chromosome.

$$cost = f(chromosome) = f(w_1, w_2, w_3, w_4) \quad (11)$$

In our case, the optimal solution is to have life-look flock behaviour. Measuring the flock behaviour can be very complicated process, expensive computationally and then time consuming. Since the purpose of this paper is to build a generic boids model and optimise it by different optimisation methods, we suggest a simple cost function. In the flowing, we explain the proposed cost function which is divided into five parts.

- Related to the alignment rule: The divergence between the direction of the boid and the average direction of the flock should be minimised. The divergence is the angle θ between the boid velocity vector \vec{v}_i and the average flock velocity vector.

$$cost_1 = \theta \quad (12)$$

- Related to the leader following rule: The divergence between the direction and the distance of the boid and the direction and the distance of the leader should be minimised.

$$cost_2 = d \quad (13)$$

Where d is the distance between p_i and p_l

$$cost_3 = \alpha \quad (14)$$

Where α is the angle between the boid velocity vector \vec{v}_i and the leader velocity vector.

- Related to the separation and cohesion rules: The boids distribution should be optimised to avoid crowding or losing contact and having nice flocking. To do this; we calculate the distance between the boid and the flock center first d_c . Then we check all the boids, if they are nearby ($<keepd$) or far enough ($>keepd$).

If ($|(\vec{p}_i - \vec{p}_j)| \leq keepd$),

$$cost_4 = \sum_{\forall b_j \in f} d_c * \frac{(|(\vec{p}_i - \vec{p}_j)| - keepd)^2}{keepd^2} \quad (15)$$

But for the far boids,

If ($|(\vec{p}_i - \vec{p}_j)| > keepd$)

$$cost_5 = \sum_{\forall b_j \in f} d_c * \frac{(|(\vec{p}_i - \vec{p}_j)| - keepd)^2}{(d_c - keepd)^2} \quad (16)$$

Then the cost is:

$$cost = cost_1 + cost_2 + cost_3 + cost_4 + cost_5 \quad (17)$$

We have used the continuous GA as explained in (Haupt, 2003). We used the parameters in (Table 1). We will analyse the results in experiment results section.

Number of optimisation variables	4
Upper limit on optimisation variables	1
Lower limit on optimisation variables	0
Maximum iteration	100
Minimum cost	0
Population size	20
Mutation rate	0.2
Selection rate	0.5

Table 1: GA parameters

PARTICLE SWARM OPTIMISATION OF BOIDS MODEL

In this section, we will give an overview of PSO algorithm in general and some examples of its applications. Then we present our proposed model (PSO for optimisation of boids model).

Particle Swarm Optimisation:

PSO is a computational method that optimises a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Kennedy and Eberhart introduced PSO in 1995 (Kennedy, 1995). PSO was originally used to solve nonlinear continuous optimization problems, but more recently it has been used in many practical, real-life application problems. For example, PSO has been successfully applied to track dynamic systems (Eberhart, 2001) and evolve weights and structure of neural networks (Zhang, 2000). PSO draws inspiration from the sociological behaviour associated with bird flocking. It is a natural observation that birds can fly in large groups with no collision for extended long distances, making use of their effort to maintain an optimum distance between themselves and their neighbours.

Cui (2009) has developed a hypried PSO and boids model (Boids-PSO), where cohesion rule and alignment

rule are both employed to improve the PSO algorithm for boids simulation and to overcome the weakness of biological background of PSO. But in our case we use PSO as an optimisation technique to optimise the coefficients in the moving vector.

The PSO methodology operates by placing a group of individual particles into a continuous search space, wherein each particle is initialised with a random position and a random initial velocity in the search space. The position and velocity are updated synchronously in each iteration of the algorithm. Each particle adjusts its velocity according to its own flight experience and the other's experience in the swarm in such a way that it accelerates towards positions that have high fitness values in previous iterations. In other words, each particle keeps track of its coordinates in the solution space that are associated with the best solution that has achieved so far by itself. This value is called personal best (*pbest*). Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighbourhood of that particle. This value is called (*best*). So the basic concept of PSO lies in accelerating each particle toward its *pbest* and the *gbest* locations, with a random weighted acceleration at each time step as shown in (figure 4).

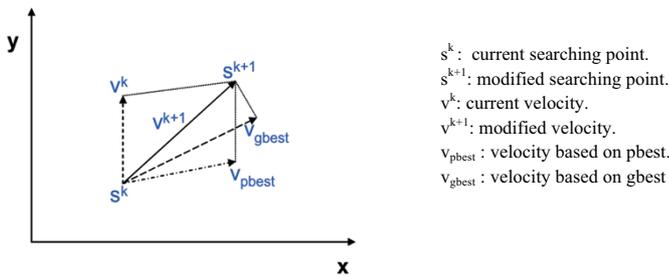


Figure 4: Concept of particle position modification by PSO

The modification of the particle's position can be mathematically modelled according to equation 18.

$$\vec{v}(k+1) = \vec{v}(k) + c_1 \vec{R}_1 (\vec{pbest} - \vec{s}_i(k)) + c_2 \vec{R}_2 (\vec{gbest} - \vec{s}_i(k)) \quad (18)$$

Where,

- $\vec{v}(k)$ is the velocity of a particle at iteration k .
- \vec{R}_1 and \vec{R}_2 are random numbers in the range of $[0,1]$ with the same size of the swarm population.
- c_1 and c_2 are learning factors which will be fixed through whole the process.

In order to improve the local search precision, Eberhart (2001) added the inertia weight w to equation 18 to be as following equation.

$$\vec{v}(k+1) = w \vec{v}(k) + c_1 \vec{R}_1 (\vec{pbest} - \vec{s}_i(k)) + c_2 \vec{R}_2 (\vec{gbest} - \vec{s}_i(k)) \quad (19)$$

The inertia weight controls the momentum of the particle by weighing the contribution of the previous velocity.

Chatterjee (2006) suggested a dynamic change of inertia weight in his work.

Clerc (1999) indicates that the use of a constriction factor K may also be necessary to ensure convergence of the particle swarm algorithm, defined as when all particles have stopped moving. Then the velocity is calculated by the equation:

$$\vec{v}(k+1) = K [\vec{v}(k) + c_1 \vec{R}_1 (\vec{pbest} - \vec{s}_i(k)) + c_2 \vec{R}_2 (\vec{gbest} - \vec{s}_i(k))] \quad (20)$$

$$K = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|} \quad (21)$$

Where $\varphi = c_1 + c_2$ and $\varphi > 4$.

Then the new position for the particles is the addition of the position at k iteration and the distance that the particles will fly with the new velocity $\vec{v}(k+1)$. The position is updated by equation 22.

$$\vec{s}_i(k+1) = \vec{s}_i(k) + \vec{v}(k+1) \quad (22)$$

The flow chart of a general PSO algorithm is shown in (figure 5).

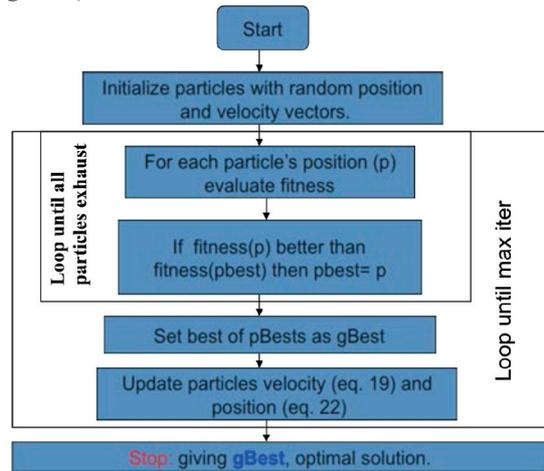


Figure 5: flow chart of general PSO algorithm

Path Planning using PSO:

One of the main applications of PSO is the path planning. PSO has been applied massively for path planning intensely in robots (Chen X 2006 and Nasrollahy 2009).

In the boids model, PSO can be applied to the flock leader, to plan and smooth his path. For this purpose; we have used PSO algorithm to plan the path to the target (i.e. food source). We have used the Euclidian distance between the particle and the target as a fitness function in the PSO.

In PSO for path planning, the inertia weight w is calculated according to the next equation.

$$w = w_{start} - \frac{w_{start} - w_{end}}{K} k \quad (23)$$

Where, κ is the iteration maximum number and k is the current iteration. By linearly decreasing the inertia weight from a relatively large value to a small value, the PSO tends to have more global search ability at the beginning of the run while having more local search ability near the end of the run.

As in robot's applications, PSO gives advantages to the path planning particularly in the dynamics environment containing stationary and moving obstacles.

We have used the parameters in (Table 2) for the PSO. In case of facing obstacles; the leader is looking to his target, if there is an obstacle whose obscures the target and the distance to this obstacle is less than a threshold, the leader will change his direction randomly to his right or his left by 45-degree angle.

Swarm size	20
Dimension of the problem	2
Maximum iteration	100
c1 (cognitive parameter)	2
c2 (social parameter)	2
C (constriction factor)	1
Inertia start	0.9
Inertia end	0.4
Upper limit on optimisation variables	100
Lower limit on optimisation variables	-100

Table 2: PSO parameters for path planning

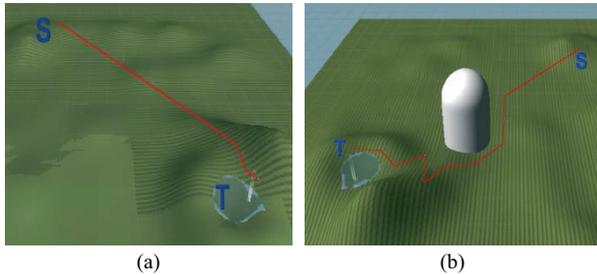


Figure 6: PSO for path planning without obstacles (a) and with an obstacle (b).

PSO for optimisation of moving vector in boids model:

As in GA for optimisation of the moving vector in the boids model, we have applied PSO to find the optimum coefficients for the moving vector (v_i) in equation 9 for each boid (b_i). We have used the same cost function as in equations from 12 to 17. We have used the PSO as explained in (Kennedy, 1995).

Population size	20
Dimension of the problem	4
Maximum iteration	100
c1 (cognitive parameter)	2
c2 (social parameter)	2
C (constriction factor)	1
Inertia start	0.9
Inertia end	0.4
Upper limit on optimisation variables	1
Lower limit on optimisation variables	0

Table 3: PSO parameters

We have used the parameters in (Table 3) and we have used equation 23 to calculate the Inertia. We will analyse the results in next section.

THE EXPERIMENT RESULTS

For testing the boids model without/with the GA and PSO, we have made a fish school in Unity3D. we have used a ready fish boid from unity website to have a nice fish shape with some animations such as moving the fish tail. The fish school has 50 fish (figure 7a). Firstly we have implemented the boids model as in equation 8 with excluding the random steering, and then we added the random steering and the leader factor. Figure 7(b,c,d) shows the simulation from different efforts. It was observed that the model need time to have a nice flocking shape with/without random movement. And the random movement was important to avoid obstacles since we didn't have a separate steering behaviour for avoiding the obstacles. The simulation of a fish school depends on the boids model as in equation 8 with equal weight coefficients gives a good flocking shape but the model was slow to get the nice flocking shape.

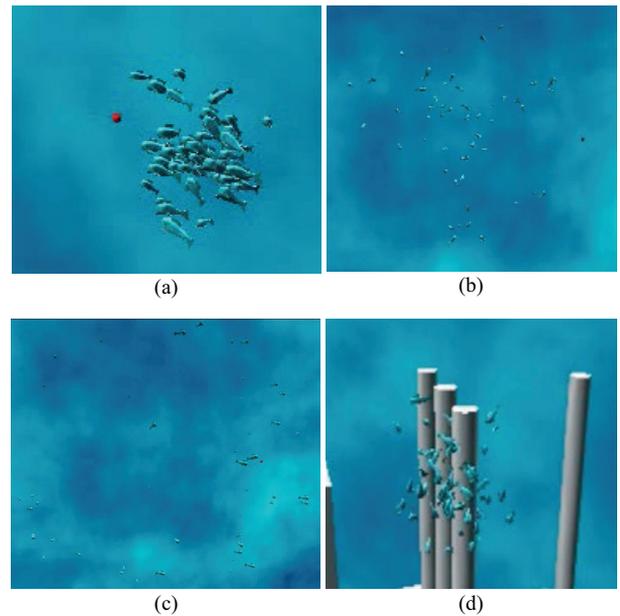


Figure 7: (a) Fish school system (b) the boids simulation without random movement after few frames from the start (c) the boids simulation with random movement (d) the boids simulation with random movement and adding leader factor and some obstacles.

Using the GA for optimisation the moving vector by finding the optimal coefficients, made the school of fish getting the nice flock shape sooner, but its very computational costly. The frame rate went down from more than 30 frames per second to almost 3 frames per second (this depends on the parameters of GA). And it is noticeable that after passing the start time and having the flock shape, there is no noticeable difference between the original boids model and the boids model

with GA. *Figure 8(a)* shows a screen shot from the simulation with GA.

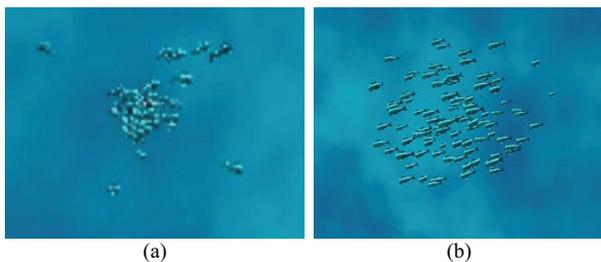


Figure 8: The boids simulation with (a) GA, and (b) PSO

PSO was faster than GA, and gave more noticeable results as shown in (*figure 8(b)*). It is observed that the boids get the flock shape faster and even the flock behaviour is look nicer than before.

We depend on our observation of the simulation results to do the comparison between the different models, because each simulation/run is different from other simulations/runs and it depends on the starting positions and the random numbers. We have selected the same population size and the number of iterations for both GA and PSO algorithms. The other parameters in GA, were selected by running the GA on many standard optimisation problems and from the literature (Haupt, 2003). The other parameters in PSO algorithm, were selected from the literature and the path planning algorithm. These parameters are selected to have a good convergence. The cost function or the objective function in general should be connected to the type of simulation. In our experiment, the objective function is to have nice fish school behaviour.

CONCLUSIONS AND FUTURE WORK

The boids model is often used in computer graphics to provide realistic life-like representations of the aggregate motion. For instance, many animations require natural-looking behaviour from a large number of characters (boids). The aggregate motion of the group is the result of the interaction of the individuals, so let each individual generate its own motion, this is easier and produces natural and unscripted motion. The individual's moving is calculated by combining all the steering behaviour vectors. To have a natural behaviour in different environments, the boid's movements should be optimised and adapted. GA and PSO algorithm are used to optimise a generic boids model by optimising the coefficients of the moving vector to minimise the cost function.

The challenge is to write rules to define natural behaviours. In the boids model, we have defined the cost function which is divided into five parts. These parts are related to the steering behaviours in the model. Thus the cost function reflects how the fish school should look in the nature. Our cost function is not computationally costly and measures simply the boids

behaviour. The cost function (objective function) should be connected to type-of-problem we want to solve, and reflects how we want the flock/swarm to behave. The setting of moving vector coefficients is determined by the cost function. We use GA and PSO to find the best coefficients values (weights of the behaviours rules) to minimise the cost function which reflects the wanted behaviour.

GA and PSO have many similarities and both of them use population-based approaches. GA is known as a good algorithm to find the global optimal solution where is PSO could stuck in the local optimum. But PSO has advantage over GA concerning the time. In the boids model, where we have adaptive boids in a dynamic environment, we are interested in a nice flocking behaviour (convergence) as in nature, and in time consuming. From the experiments for both GA and PSO, we observed that PSO is much faster than GA, and gives a faster convergence, and because the PSO is computationally less costly than GA, we could notice the convergence more in PSO than GA.

The challenge is to balance between the convergence (nice flocking behaviour and the adaptively) and the time consuming. Having more advanced cost function probably will give better results, but it will be very expensive and lead to a very slow model. Applying the optimisation algorithm not continuously such as applying the optimisation algorithms for only some parts of the simulation such as at beginning of the simulation until the boids get a nice flock shape which is wanted, or when the boids facing obstacles or enemies, will accelerate the model.

REFERENCES

- Anthony L. (2002), "Artificial life", Macmillan Press Ltd., Basingstoke, UK.
- Chatterjee A., Siarry P. (2006), "Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimisation", *Computer & Operations Research* 33, 859-871.
- Chen, Y.-W.; Kobayashi, K.; Huang, X. & Nakao, Z. (2006), "Genetic Algorithms for Optimization of Boids Model", in Bogdan Gabrys; Robert J. Howlett & Lakhmi C. Jain, ed., 'KES (2)', Springer, , pp. 55-62 .
- Chen X.; Yangmin L. (2006), "Smooth Path Planning of a Mobile Robot Using Stochastic Particle Swarm Optimization", *Mechatronics and Automation, Proceedings of the 2006 IEEE International Conference on* , vol., no., pp.1722,1727, 25-28.
- Clerc M. (1999), "The swarm and the queen: towards a deterministic and adaptive particle swarm optimization", in *Proceedings of the Congress of Evolutionary Computation*, Washington, DC, pp. 1951-1957.
- Cui Z., Shi Z. (2009), "Boid particle swarm optimisation", *International Journal of Innovative Computing and Applications* 2 (2): 77-85.
- Delgado M. C., Ibanez J., Bee S., et al. (2007), "On the use of Virtual Animals with Artificial Fear in Virtual Environments", *New Generation Computing* 25 (2): 145-169.
- Docs.unity3d.com (2014), *Unity - Getting started with Mono Develop*. [ONLINE] Available at:

- <http://docs.unity3d.com/Documentation/Manual/HOWTO-MonoDevelop.html>. [Accessed 31 January 2014].
- Eberhart R., Shi Y. (2001), "Tracking and optimizing dynamic systems with particle swarms", Proc. Congress on Evolutionary Computation 2001, Seoul, Korea
- Fogel D. (2000), "Evolutionary Computation: Towards a New Philosophy of Machine Intelligence", IEEE Press, New York.
- Hartman C., Benes B. (2006), "Autonomous boids", *Computer Animation and Virtual Worlds* 17 (3-4): 199–206.
- Haupt R., Haupt S. (2003), "Practical Genetic Algorithms", 2nd Ed, Wiley, 2003.
- Holland J. (1975), "Adaptation in Natural and Artificial Systems", Ann Arbor: University of Michigan Press.
- Ice', S. and Malone, L. (2014), "*Stanley and Stella in 'Breaking the Ice' (1987)*", [online] Available at: <http://www.imdb.com/title/tt0302371/> [Accessed: 31 Jan 2014].
- Kennedy J., Eberhart R. (1995), "Particle Swarm Optimisation", *Proceedings of IEEE International Conference on Neural Networks IV*. pp. 1942–1948.
- Nasrollahy, A.Z.; Javadi, H. (2009), "Using Particle Swarm Optimization for Robot Path Planning in Dynamic Environments with Moving Obstacles and Target", *Computer Modeling and Simulation, 2009. EMS '09. Third UKSim European Symposium on*, vol., no., pp.60,65, 25-27.
- Returns, B., Burton, T., Kane, B., Waters, D., Keaton, M., Devito, D. and Pfeiffer, M. (2014), "Batman Returns (1992)" *IMDb*, [online] Available at: <http://www.imdb.com/title/tt0103776/> [Accessed: 31 Jan 2014].
- Reynolds C. (1987), "Flocks, Herds, and Schools: A Distributed Behavioural Model", *Computer Graphics*, 21:4,1987, 25-34.
- Reynolds C. (1999), "Steering behaviour for autonomous characters", <http://www.red3d.com/cwr/steer/>, first version from 1999.
- Unity3d.com (2014), *Unity - Game Engine*. [ONLINE] Available at: <http://www.unity3d.com>. [Accessed 31 January 2014].
- Valvesoftware.com (2014), *Valve*. [ONLINE] Available at: <http://www.valvesoftware.com/>. [Accessed 31 January 2014].
- Zhang C., Shao H., Li Y. (2000), "Particle Swarm Optimisation for Evolving Artificial Neural Network", In the 2000 IEEE International Conference on Systems, Man, and Cybernetics, vol.4, pp.2487-2490.

AUTHOR BIOGRAPHIES

SALEH ALALIYAT was born in Jenin, Palestine. He is currently working as a PhD candidate at Aalesund University College, Norway. He received his Master's degree in Media Technology from Gjøvik University College in Norway.

HARALD YNDESTAD was born in Aalesund, Norway. He has studied cybernetics at the University in Trondheim, obtained a dr.philos degree in 2004 and he is now professor at Aalesund University College. His research interests are complex systems, swarm intelligence and ecosystem dynamics.

FILIPPO SANFILIPPO is a PhD candidate in Engineering Cybernetics at the Norwegian University of Science and Technology, and a research assistant at the Department of Maritime Technology and Operations, Aalesund University College, Norway. He obtained his Master's Degree in Computer Engineering at University of Siena, Italy.

MODELLING OF PHOTOVOLTAIC ENERGY GENERATION SYSTEMS

Pekka Ruuska
Antti Aikala
Robert Weiss

VTT Technical Research Centre of Finland
P.O.Box 10000, FI-02044 VTT, Finland
E-mail: pekka.ruuska@vtt.fi

KEYWORDS

APROS, Dynamic simulation, Modelling, Photovoltaic, Solar power.

ABSTRACT

We present the development of a dynamic software simulation environment for modelling small and large scale photovoltaic (PV) energy generation systems. The model elements of solar panels, Maximum Power Point Trackers (MPPT) and solar inverters were integrated to a simulator system providing weather and power grid models as well as simulation of combustion power plants. The integrated system enables accurate dynamic simulation of complicated energy processes that utilize a variety of technologies. We tested our model against a large scale solar photovoltaic plant that produces energy to the power grid. The first test runs proved that the implemented models predict the produced energy adequately under constant irradiation.

INTRODUCTION

Simulation can help to maximize the generated photovoltaic energy and to optimize the design of wide solar panel systems. Furthermore, the simulation models can be integrated with models of other energy generation systems such as heat and power plants. Dynamic simulations provide tools for aiding the design work, optimization, accident analysis and control development of power generating systems. They can advance energy-efficiency, reduce greenhouse gas emissions and cut costs of producing energy in complicated industry processes.

We developed a software environment for modelling photovoltaic panels, maximum power point trackers (MPPT) and solar inverters. These were integrated to APROS, which is a widely used dynamic process simulator. APROS provides weather models, power grid simulation as well as detailed and accurate modelling of combustion power plants.

This paper describes calculation principles, design and implementation of the new models. First test runs and

comparison to a real PV energy generation system is presented. These simulations were done in the EuroEnergest project, funded by the European Union's 7th Framework Programme. Among the goals of that project is to reduce energy consumption in car industry through exact and predictive control of on-site energy production processes.

THE APROS ENVIRONMENT

APROS offers tools, solution algorithms, and model libraries for the full-scale modelling and simulation of dynamic processes. APROS was developed by VTT Technical Research Centre of Finland and Fortum Corporation. It can be utilized e.g. in the transient modelling of various nuclear and combustion power plants, pulp and paper mills, fuel cells and several other energy systems. The process itself, as well as automation and electrical systems can be modelled, and real plant measurements can be implemented to APROS (Saarinen et al. 2007). APROS is currently developed for modelling the alternative and small-scale energy production systems.

MODELLING PHOTOVOLTAIC PANELS

As an electrical element a photovoltaic panel resembles a DC current source still its internal resistance is not constant; it varies non-linearly with solar irradiation. Typically the current produced by photovoltaic modules depends on irradiance and cell temperature while the voltage depends mainly on temperature (Figure 1).

Precise mathematical modelling of a PV cell requires determining physical characteristics, which are usually not provided in the solar panels' data sheets (Gow and Manning 1999). Strict models can also require utilizing complicated numerical methods in calculations. A solar cell can be modelled with a single diode model (Figure 2). At low irradiance a two diode model should yield more precise results. An additional parallel resistance may further advance the accuracy of the model at the expense of more complex calculations. In all models the mathematical difficulties arise from the p-n diode junction of the equivalent circuit as it must be analysed with an exponential equation (Sera et al. 2007).

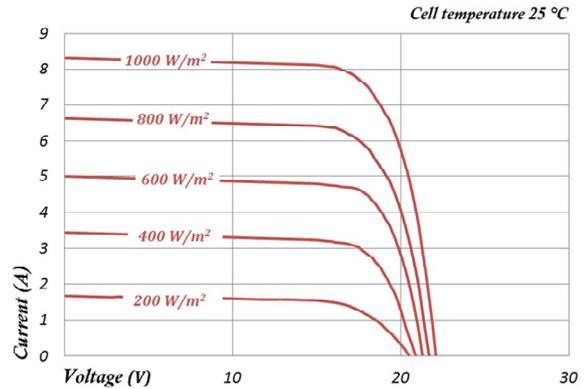
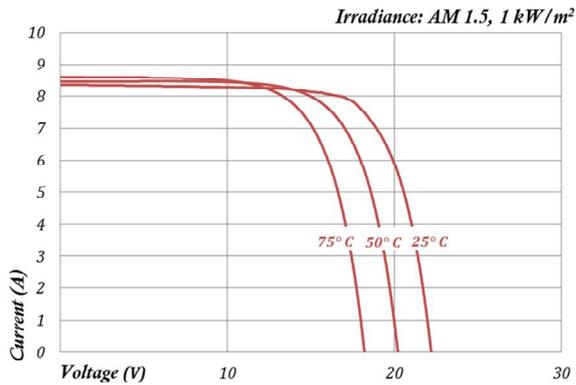
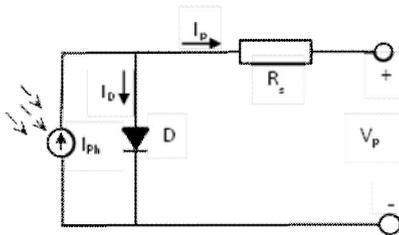


Figure 1: V-I Curve for A Typical Multicrystal Photovoltaic Module under 1,0 kW/m² Irradiation in Various Temperatures (on the left) and The Same Cell at T = 25° C under Different Radiation Intensities (Kyocera 2013)

However, as the electrical parameters of real solar modules always differ somewhat from the theoretical values, quite precise results from simulations cannot be expected. Also the electrical characteristics and the performance of solar panels slowly vary through their lifetime. The changes caused by ageing and soiling can be exactly monitored only through measuring. Furthermore, cloudiness and other weather conditions crucially affect the performance of a solar panel system and these factors are also difficult to predict or measure exactly. Therefore simplified models which do not produce quite accurate results in all conditions but which do not entail such impractical requirements, can be quite appropriate in simulation.



Figures 2: A Simplified Single-Diode Equivalent Circuit of a PV Module (Xiao et al. 2004; Bellini et al. 2009)

We implemented two different mathematical models to our simulation software. With the first model we can determine the output current and voltage under varying conditions from a solar panel's data sheet information (Bellini et al. 2009). We refer to this model as "Model 1" in this paper. The second model ("Model 2") is more accurate especially under low insolation, but it requires first measuring I_{sc} (Short Circuit Current) and V_{oc} (Open Circuit Voltage) under two different radiation intensities (Zhou et al. 2007). As we first ran the simpler model in our simulations and experiments, we do not describe in detail the second model in this paper.

A simplified linear formula for the I_{sc} of a photovoltaic cell is (Bellini et al. 2009):

$$I_{sc} = (G/G_0) \cdot [I_{SC0} + k_T(T - T_0)] \quad (1)$$

where k_T is temperature coefficient for I_{sc} (in A/°C), G is solar irradiation and T is cell temperature while G₀, I_{SC0} and T₀ are the corresponding values under STC (Standard Test Conditions). The panel manufacturers provide temperature coefficients for I_{sc} and V_{oc} in their data sheet. The photovoltaic current I_p as a function of photovoltaic voltage V_p is determined by:

$$I_p = I_{SC} [1 - C_1(e^{(V_p/C_2 \cdot V_{OC})} - 1)] \quad (2)$$

Where coefficients C₁ and C₂ are defined in Equations (3) and (4) and which depend on the parameters I_{SC0}, V_{OC0}, I_{MPP0} (Current at Maximum Power Point) and V_{MPP0} (Voltage at Maximum Power Point). Also these parameters are always presented in data sheets of PV panels.

$$C_1 = (1 - I_{MPP0} / I_{SC0}) \cdot e^{(-V_{MPP0}/C_2 \cdot V_{OC0})} \quad (3)$$

$$C_2 = ((V_{MPP0} / V_{OC0}) - 1) / \ln(1 - (I_{MPP0} / I_{SC0})) \quad (4)$$

It is shown in (Bellini et al. 2009) that from the above we get V_{oc} under radiation intensity G and in temperature T as follows:

$$V_{OC} = V_{OC0} + m_T(T - T_0) - \Delta V(G) \quad (5)$$

where m_T is the temperature coefficient for V_{oc} (in V/°C) and ΔV(G) is a correction term obtained from:

$$\Delta V(G) = V_{OC0} - V_{OCm} \quad (6)$$

in which V_{OCm} is the transposed open circuit voltage derived from yet another parameter I_t(G) (transposed current) as follows:

$$I_t(G) = I_{SC0} \cdot [1 - (G/G_0)] \quad (7)$$

$$V_{OCm} = C_2 \cdot V_{OC0} \cdot \ln[1 + (1 - (I_t(G)/I_{SC0})) / C_1] \quad (8)$$

where G_0 is 1000 W/m^2 while C_1 and C_2 are defined in equations (3) and (4).

MODELLING MAXIMUM POWER POINT TRACKERS

Maximum power point trackers (MPPT) are used to maximize the output power from solar panels. The goal of the MPPT techniques is to automatically find the voltage V_{MPP} or current I_{MPP} at which the PV array (solar panel) should operate to obtain the maximum power output P_{MPP} under a given temperature and irradiance. Even a simple MPPT may yield an energy gain of 20% or up to 30% from a PV array; therefore they have become essential elements in PV systems today. An MPPT can be integrated to a solar inverter or to a battery charger or they can be installed as a separate element into a PV system.

Dozens of different algorithms to maximum power point tracking have been introduced (Esram and Chapman 2007). These can be implemented with various technologies. The algorithms perform differently when irradiation changes, some of them are better at low light and others may find the MPP quicker in steady conditions. Still the MPPT vendors usually do not publish the algorithms which they implement in their devices. Therefore we can present only approximations of the real characteristics of the MPPT devices that we simulate. However, already our first test runs indicated that our estimations were rather close to reality.

We chose two different algorithms to implement into our simulation model. These are the Incremental Conductance Method and the Fractional Open Circuit Voltage Method. Most of the newer and perhaps more sophisticated algorithms are based on these rather classic approaches. The first method is supposed to determine the MPP optimally and without oscillations while it may track in the wrong direction in some conditions. The second method never finds the exact MPP; still it is shown that the method works better than many others under low insolation (Esram and Chapman 2007).

In the first procedure the output voltage V is changed in steps ΔV and the resulting changes of output voltage and current I are measured. If the change in conductance $\Delta G = \Delta I / \Delta V$ is positive the output voltage is increased again until the change is zero. And if the first change in conductance is negative, the voltage is decreased until there is no change in conductance.

The second method is very simple, it requires only measuring the output voltage and factoring that with a constant. The constant is known (0.76 is typical) or it can be determined by measuring.

TEST RUNS AND RESULTS

Our first example illustrates generation of DC power with a photovoltaic system (Figure 3). On the left in the figure are the solar radiation and the irradiation processor modules that calculate the total amount of effective received irradiation on a tilted surface. The output voltage of the PV panel is controlled by the MPPT module which changes the voltage ratio of the DC/DC converter. The model calculates the produced power. The generated electricity is supplied to a battery or a DC load module.

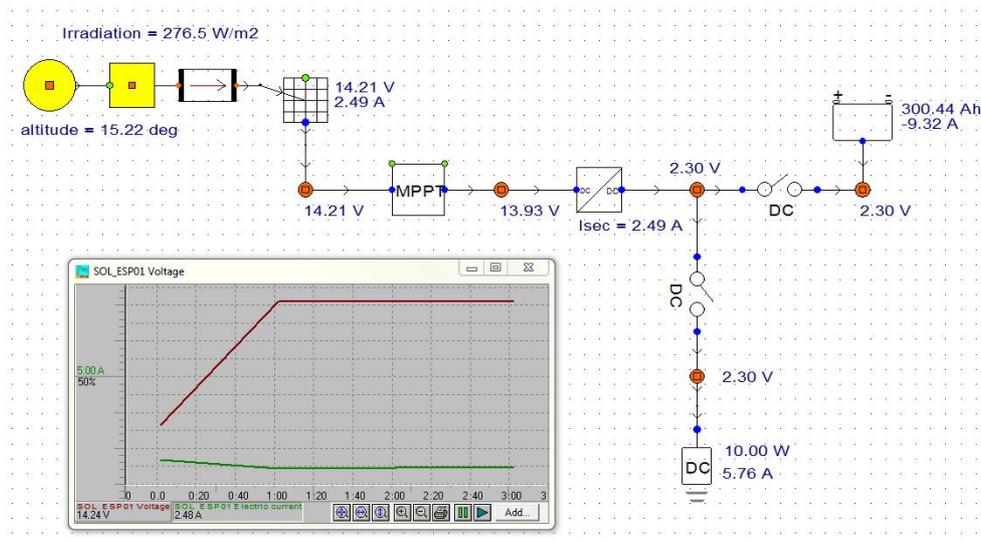


Figure 3: APROS Simulation of a Small-scale PV Energy Generation System

Examples of I-V curves calculated by using “Model 1” and the system in Figure 3 are depicted in Figure 4. It shows that at radiation intensity of 400 W/m² or more the discrepancy with the data sheet (Figure 1) curves is less than 2 % while at low irradiation it is close to 10%.

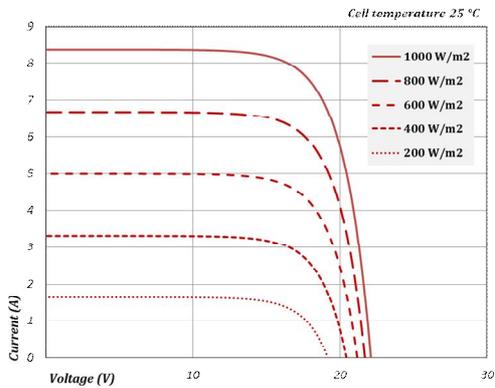


Figure 4: Calculated V-I Characteristics at Different Irradiancies with “Model 1”

After this we simulated solar PV energy production to the power grid by connecting solar panel modules in series with a MPPT, a DC/DC converter and an inverter to build up a “solar inverter” (Figure 5). In these simulations we utilized “Model 1”

and the Incremental Conductance Method in MPP tracking (Esram and Chapman 2007). The reason for these selections was that we wanted to analyse the energy production potential of the system under high irradiation, when it is connected to the grid. “Model 1” is adequate for that purpose and it needs only the data sheet information as input, which facilitates the simulation.

In the first simulation run we kept the weather model’s parameters constant to analyse the accuracy of our photovoltaic cell, MPPT and solar inverter models (Figure 6). The tests indicated that our simulations can give rather close estimation of the energy produced at the site, if the weather conditions remain steady. The error between the measured and the calculated energy was about 2 % when the Sun was high and the local cloud conditions were consistent (Table 1). The cloudiness parameter of the APROS weather model was set to 0.45 indicating “partly cloudy” sky. Such cloudiness prevailed only between 10:30 and 11:30 on that day. From 11:30 to 14:00 the difference between the expected and measured energy varies from 3% to 10%. After 14:15 and up to 17:00 the discrepancy is about 6%. The inaccuracy is mostly explained by the changes in cloudiness. Another important factor is the low resolution of the measuring system; it provides the produced energy as kWh in 15 minutes which averages the peaks in produced photovoltaic energy.

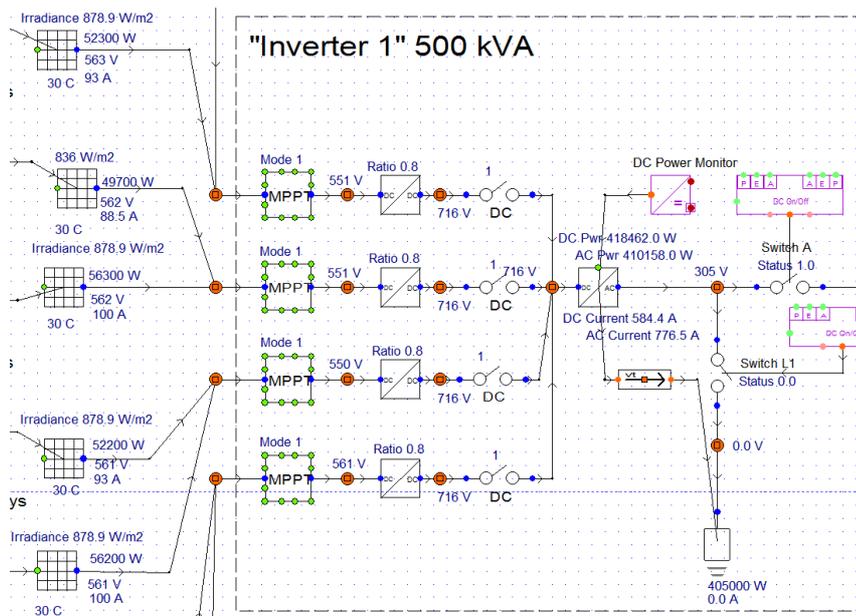


Figure 5: Part of a Model that Simulates a PV Production System Capable to 500 kW of AC Power.

In the morning hours, before 09:00 the sky was mostly clear and setting the cloudiness to 0.45 gave too low output voltage. There are also two further reasons for the clearly bigger differences between the measured and the calculated energy in the low light conditions. Firstly, the simple mathematical model employed in these tests may give a 10% error in those conditions (Bellini et al. 2009). Secondly, when the Sun is low on the horizon, the PV panels generate so low voltages (less than 500 V) that the solar inverters cannot be connected to the power grid. Therefore, when no load is connected, the simulation models of the electrical components do not calculate worthy results either.

Table 1: Comparison of Calculated and Measured Energy as kWh in 15 Minutes

Time	Measured	Calculated	Difference
10:30	296	290	-2.1 %
10:45	312	313	+0.2 %
11:00	329	334	+1.5 %
11:15	345	354	+2.5 %
11:30	363	372	+2.5 %
11:45	376	389	+3,4 %

In our second tests, publicly available weather information from Martorell was utilized. The cloudiness was mapped to the weather model as numbers between 0.10 and 0.55. In this case, the calculated energy differs from the measured production with an error varying from 2% to 7% (Figure 7).

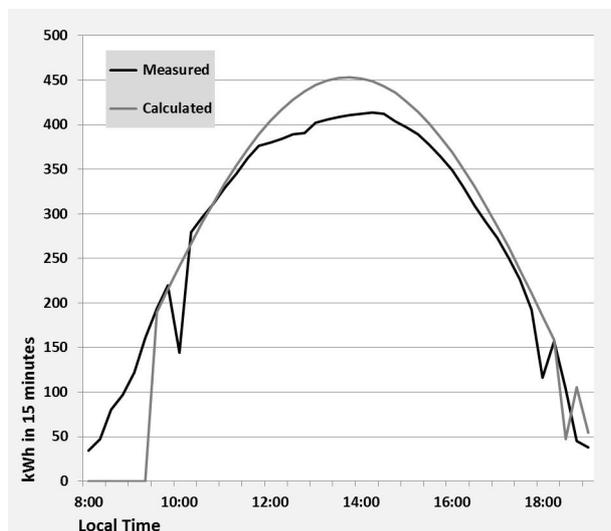


Figure 6: Comparison of the Measured and the Simulated Energy Produced on 10th July, 2013

Unfortunately, typical weather data describes cloudiness only as words; such as “scattered clouds”. And this information is given as average for 30 minute periods, while the received irradiation may rapidly vary on a wide scale. It is clear that the 30 minute intervals

are too long for our tests (Hansen et al. 2102). In our example case, the biggest error is at 13:30, when the cloudiness changes from “partly cloudy” to “scattered clouds”. And there are other environmental factors such as local geography and site topology that affect the received irradiation. Quantified irradiation data can be provided with a local pyranometer, which we plan to utilize in the next phase of our project. Still we cannot expect that the accuracy of our models would radically advance as there are several uncertainties in the model. The total inaccuracy of our simulation is typical when compared to other PV energy yield simulations (Ransome 2008).

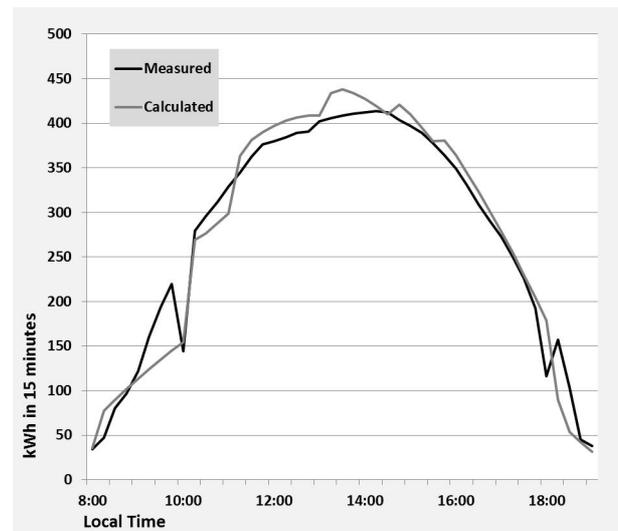


Figure 7: The Same Comparison as in Figure 6 when also the Changes in Cloudiness are simulated

CONCLUSIONS AND FURTHER WORK

Simulation models of PV cells, MPPT and solar inverters were implemented into APROS system and the first calculation results were compared against a real large-scale PV plant. The mathematical models were shown to produce adequately accurate results. However, more precise results can be expected, when more exact quantitative information about the weather conditions become available. One of the aims of the EuroEnergest project is to accurately forecast the total energy generation potential of the pilot plant. This requires modelling at least the solar PV systems, the on-site CHP plant, the boiler systems, the heat recovery system and the absorption chillers with APROS. The implemented model of the solar PV energy systems is sufficient for that purpose.

REFERENCES

- Bellini A.; S. Bifaretti.; V. Iacovone and C. Cornaro. 2009. “Simplified Model of A Photovoltaic Module”, in *2009 Applied Electronics*, (Pilsen, Czech Republic, Sep 9-10), IEEE AE 2009, 40-45

- Esrasm T. and P. L. Chapman. 2007. "Comparison of Photovoltaic Array Maximum Power Point Tracking Techniques" *IEEE Transactions on Energy Conversion* 22, No.2 (Jun), 439-449
- Gow A. and C.D.Manning. 1999. "Development of A Photovoltaic Array Model for Use In Power-Electronics Simulation Studies", *IEE Proceedings Electric Power Applications* 146, Issue 2 (Mar)
- Hansen C. W.; J. S. Stein and D. Riley. 2012. "Effect of Time Scale on Analysis of PV System Performance" *Sandia Report, SAND2012-1099*, Sandia National Laboratories, Livermore, California, USA
- Kyocera. 2013. Kyocera Fineceramics GmbH Solar Division, "Kyocera KD135SX-1PU High Efficiency Multicrystal Photovoltaic Module", *Datasheet information, Kyocera Solar*, Esslingen, Germany
- Ransome S. 2008. "Modelling inaccuracies of PV energy yield simulations", in *Proceedings of the 33rd PV Specialists Conference*, (San Diego, CA, USA, 2008)
- Saarinen, J.; M. Halinen; J. Ylijoki; M. Noponen; P. Simell and J. Kiviaho. 2007. "Dynamic Model of 5 kW SOFC CHP Test Station." *Journal of Fuel Cell Science and Technology*, 4, Issue 4 (Nov), 397-405
- Sera D.; R. Teodorescu and P. Rodriguez. 2007. "PV Panel Model Based on Datasheet Values", in *IEEE International Symposium on Industrial Electronics, ISIE 2007*. (Vigo, Spain, Jun 4-7) Electrical Engineering/Electronics, Computer, Communications and Information Technology Association, 2392-2396.
- Xiao W.; W. G. Dunford and A. Capel. 2004. "A Novel Modeling Method for Photovoltaic Cells", in *Power Electronics Specialists Conference, 2004* (Aachen, Germany, Jun 2004), IEEE PESC 04, 1177-1183
- Zhou W.; H. Yang and Z. Fang. 2007. "A Novel Model for Photovoltaic Array Performance Prediction", *Applied Energy* 84 (2007), *Applied Energy* 84, Elsevier, (Nov) 1187-1198

AUTHOR BIOGRAPHIES

PEKKA RUUSKA received his M. Sc. degree in Electrical Engineering in 1980 and Licentiate of Technology (Computer Technology) in 1993 from the University of Oulu, Finland. From 1985 he has been

working with VTT Technical Research Centre of Finland. He is currently a senior scientist and his research interests are energy efficiency and renewable energy systems. His e-mail address is: Pekka.Ruuska@vtt.fi and his Web-page can be found at <http://www.vtt.fi/?land=en>

ANTTI AIKALA received his M. Sc. degree in Forest Products Technology, in 2000 from the Helsinki University of Technology. He has worked as a researcher first in Keskuslaboratorio Oy, a research company for Forest industries in Finland, and from 2009 as a research scientist in VTT Technical Research Centre of Finland. He has worked in areas of simulation, statistical analysis and optimization in the field of dynamic systems and processes. His e-mail address is: antti.aikala@vtt.fi and his Web-page can be found at <http://www.vtt.fi/?land=en>

ROBERT WEISS was born in Nürnberg, Germany and went to Helsinki University of Technology, Finland, where he studied engineering physics and obtained his M.Sc. (Tech.) degree in 1993 and Licentiate (Tech.) degree in 2002. Since 1990 he has worked with energy and environment field with research, IT, simulation and management topics for Energy Consultancy companies, VTT Technical Research Centre of Finland, ABB Ltd, and Process Vision Oy. Since 2010 he is managing and coordinating Smart Energy Grid related research and simulation projects at VTT Technical Research Centre of Finland. His email address is: Robert.Weiss@vtt.fi and his Web-page can be found at <http://www.vtt.fi/?land=en>

Genetic algorithm with simulation for scheduling of a flow shop with simultaneously loaded stations

Professor Dr.-Ing. Frank Herrmann

Ostbayerische Technische Hochschule Regensburg – Technical University of Applied Sciences Regensburg

Innovation and Competence Centre for Production Logistics and Factory Planning (IPF)

PO box 120327, 93025 Regensburg, Germany

E-Mail: Frank.Herrmann@OTH-Regensburg.de

KEYWORDS

Simulation of restrictions, scheduling, flow-shop, no-buffer (blocking), no-wait, genetic algorithm, real world application.

ABSTRACT

In this study, a real world flow shop with a transportation restriction is regarded. This restriction reduces the set of feasible schedules even more than the no-buffer restrictions discussed in the literature in the case of limited storage. Still this problem is NP-hard. Since this scheduling problem is integrated in the usual hierarchical planning, the tardiness is minimised. Compared to even specific priority rule for this class of problems the suggested genetic algorithm delivers significant better results. The specific structure of this class of problems complicates the calculation of the performance criteria. This is solved by a simulation algorithm.

1. INTRODUCTION

Specific products are produced by special machines which are often grouped in a flow shop. They have to produce small batches with short response times, so scheduling algorithms are needed to ensure that under the constraint of a high average load of the flow shop, the due dates of the production orders are met. Nowadays, such special designed flow shops often have technological restrictions, which complicate the scheduling. For example in cell manufacturing, buffer could be non-existent due to limited space and storage facilities. So, in recent years, a considerable amount of interest has arisen in no-buffer (blocking) scheduling problems and in no-wait scheduling problems, with makespan as objective criteria. Often these production systems deliver products for other systems as well. Due to the hierarchical planning which is implemented in enterprise resource planning systems (ERP system) (see e.g. Jacobs et al. 2010), the local completion times in one production system in many cases determine the

earliest possible starting times in another production system. Thus, the delay of the operations in a production system has an impact on the effectivity of this coordination process. Therefore, tardiness is considered as objective criteria.

2. A REAL WORLD APPLICATION

The problem is a modification of a partly automated production line at Fiedler Andritz in Regensburg to produce filter (baskets) with a lot size of 1. All filters have unified constructions. They differ in varying heights of the baskets and there exist different designs. The production line consists of 4 stations which are shown in Figure 1. Station 1 assembles 6 single batons (called consoles) on an assembly ground plate to a skeleton of a filter basket. Baton profiles are assembled into the provided slots of the filter basket skeletons. At the plunge station a wire coil is contrived in the device of a lining machine. The lining machine straightens the wire and inserts batons into the slots. To ensure stability, the span station installs span kernels in the case of outflow filter baskets and span belts in the case of inflow filter baskets. Then, the filter basket is lifted from the assembly ground plate and is transported to the welding station, at which the baton profiles are welded on the filter basket skeletons. The accomplished filter basket leaves the production line. Prior to this, the span medium is removed. An overhead travelling crane lifts a filter basket out of a station, transports it to the next station and inserts it directly in this station. This is just possible if this station is free. So, there is no buffer in the production line and each feasible schedule of jobs is a permutation of these jobs. Due to other operational issues the crane can just be moved if all stations are inactive. Since an operation cannot be interrupted, the transport has to be performed after the completion of all operations on the stations in the flow shop. Due to further operational issues this restriction has to be applied also for the first and the last station; note, that the crane loads S1 and unloads S4 as well. In summary,

all stations are loaded and unloaded with filters during a common process and this process starts with the last station S4, followed by station S3, S2, until station S1 is reached. It is allowed that a station is empty; then this station is skipped (may be partially) in this process.

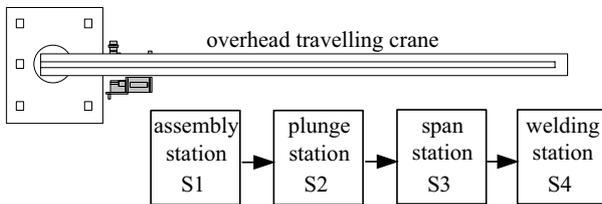


Figure 1: Structure of the production line

There are 10 part types. Their different heights or designs causes different processing times which are listed in Table 1.

Table 1: Processing times for all part types in minutes

Part type	Station				Sum of times
	S1	S2	S3	S4	
P1	100,5	50	53,5	9	213
P2	256,5	50	53,5	9	369
P3	122	135	90	75	422
P4	256,5	50	267	9	582,5
P5	182	200	135,5	140	657,5
P6	100,5	300	53,5	300	754
P7	223	250	196	220	889
P8	223	250	206,5	220	899,5
P9	100,5	300	267	300	967,5
P10	256,5	300	267	300	1123,5

At the company's production site, the jobs for filters are generated by an SAP system and produced filters are stored before they are assembled into other products or sold directly to customers. Therefore, all jobs with a release date after the beginning of a period are released at the beginning of this period. One period consists of one day with three 8 hour shifts. For this investigation, sequences of jobs of filter types with lot size 1 are randomly generated for each period t by an generating algorithm which has been designed in accordance with the proceeding in (Russel et al. 1987) and in (Engell et al. 1994): An additional filter type F , released in period t , consumes capacity on each station during the time between t and its due date; the calculation for the capacity just uses the net processing time and does not regard the dependencies between the jobs (released so far). F is accepted as long as this consumed capacity on each station is below a maximal load level, otherwise it will be skipped to the next period. A maximal load level is an (intended) average load ($L_0(S)$) plus 0, -30% and +30% of $L_0(S)$. Over the first 5, 10, 15 etc.

consecutive periods, the load level variations average to zero.

In reality at the company, there are large numbers of periods with a low number of late jobs and large numbers of periods with a high (or even very high) number of late jobs. To achieve a comparable situation

for this investigations, due dates are determined in a way so that scheduling with the FIFO rule (first-in-first-out) causes a specific percentage of late jobs. The company confirmed that job sequences with 30%, 50%, 70% and 85% of late jobs by scheduling with the FIFO rule (called time pressure) are comparable to the ones which occurred in the real operation. As a result of the generating algorithm's calculations the mean difference between the due dates and the release dates are between 2 and 3,5 days with a standard deviation of 0,5 days. Andritz Fiedler has confirmed that such timeframes for processing jobs are representative.

The time needed for loading and unloading a station is negligible compared to the duration of the operation itself. In addition this task is independent from the allocation (or loading) of other stations and the required time is included in the duration of the operation.

The general scheduling problem consists of M stations and a pool of N jobs, which may change at any time, with known earliest possible starting times for release dates a_i ($1 \leq i \leq N$) and due dates f_i ($1 \leq i \leq N$) respectively. Also there is the duration $t_{i,j}$ of operation ($o_{i,j}$) j ($1 \leq j \leq M$) of job i ($1 \leq i \leq N$), which is being worked on station j . Due to the reasons, said in the introduction, as performance criteria average tardiness (T_{Mean}) and standard deviation of the tardiness (T_{σ}) are primarily analysed.

The time between two consecutive executions of the load process is determined by the maximum of the duration of the operations (including setup time) on the stations in the flow shop. This is called cycle time. This "load"-restriction, the no-buffer condition and the capacity of the stations are the main restrictions.

The no-buffer condition means a relaxation of the scheduling problem with the (above) "load"-restriction. Scheduling problems with the no-buffer are proven to be NP-hard in the strong sense for more than two stations; see e.g. (Hall and Sriskandarajah 1996), which contains a good survey of such problems.

3. LITERATURE REVIEW

As mentioned earlier, the real application is close to the class of no-buffer (blocking) scheduling problems. Solutions for the no-buffer (blocking) scheduling problems are published in various papers. Most papers minimise makespan as the ones in the following review – later publications to minimise tardiness are reviewed. Thus, the following review explains a large spectrum of possible procedures. In (McCormick et al. 1989) a schedule is extended by a (unscheduled) job that leads to the minimum sum of blocking times on machines which is called profile fitting (PF). Often the starting point of an algorithm is the NEH algorithm presented in (Nawaz et al. 1983), as it is the best constructive heuristic to minimize the makespan in the flow shop with blocking according to many papers, e.g. (Framinan et al. 2003). Therefore, (Ronconi 2004) substituted the

initial solution for the enumeration procedure of the NEH algorithm by a heuristic based on a makespan property proven in (Ronconi and Armentano 2001) as well as by the profile fitting (PF) of (McCormick et al. 1989). (Ronconi 2005) used an elaborated lower bound to realise a branch-and-bound algorithm which becomes a heuristic since the CPU time of a run is limited. Also for minimising makespan, (Grabowski and Pempera 2007) realised and analysed a tabu search algorithm. As an alternative approach, (Wang and Tang 2012) have developed a discrete particle swarm optimisation algorithm. In order to diversify the population, a random velocity perturbation for each particle is integrated according to a probability controlled by the diversity of the current population. Again, based on the NEH algorithm, (Wang et al. 2011) described a harmony search algorithm. First, the jobs (i.e. a harmony vector) are ordered by their non-increasing position value in the harmony vector, called largest position value, to obtain a job permutation. A new NEH heuristic is developed on the reasonable premise that the jobs with less total processing times should be given higher priorities for the blocking flow shop scheduling with makespan criterion. This leads to an initial solution with higher quality. With special settings as a result of the mechanism of a harmony search algorithm, better results are achieved. Also (Ribas et al. 2011) presented an improved NEH-based heuristic and uses this as the initial solution procedure for their iterated greedy algorithm. A modified simulated annealing algorithm with a local search procedure is proposed by (Wang et al. 2012). For this, an approximate solution is generated using a priority rule specific to the nature of the blocking and a variant of the NEH-insert procedure. Again, based on the profile fitting (PF) approach of (McCormick et al. 1989), (Pan and Wang 2012) addressed two simple constructive heuristics. Then, both heuristics and the profile fitting are combined with the NEH heuristic to three improved constructive heuristics. Their solutions are further improved by an insertion-based local search method. The resulting three composite heuristics are tested on the well-known flow shop benchmark of (Taillard 1993), which is widely used as benchmark in the literature.

To the best of my knowledge, only a few studies investigate algorithms for the total tardiness objective (for flow shops with blocking). (Ronconi and Armentano 2001) have developed a lower bound which reduces the number of nodes in a branch-and-bound algorithm significantly. (Ronconi and Henriques 2009) described several versions of a local search. First, with the NEH algorithm, they explore specific characteristics of the problem. A more comprehensive local search is developed by a GRASP based (greedy randomized adaptive search procedure) search heuristic. There are just a few genetic algorithms with this performance criteria: for a no-wait flowshop scheduling problem one is published in (Chaudhry and Mahmood 2012) and for a flowshop with blocking one is published in (Januario et al. 2009).

3. GENETIC ALGORITHM

In the literature to scheduling problems each genetic algorithm has typically the following basic structure (s. e.g. (Werner 2013)):

0. Representation of a schedule by a chromosome
 1. Initial population
 2. Fitness of a (actual) population
 3. Selection
 4. Application of a genetic operation
 5. Formation of the new population
 6. If stopping criteria is not satisfied, then go to step 2
 7. Selection of a best chromosome (i.e. schedule)

Due to the load condition each feasible schedule is a permutation of jobs. So, this permutation can be used as a chromosome. Above all, the load restriction determines the allocation of each operation on a station during the execution of the (above) permutation of the jobs. This is simulated by an algorithm. So, a performance criteria as tardiness can be calculated. (Note: This concept can be extended to further restrictions – which are not covered by the representation of a schedule.)

Initial populations are generated either by accident or by a priority rule as well as standard heuristics for flow shop problems with a relatively low runtime. Implemented are many rules and concrete settings are said in chapter “experimental calibration of the genetic algorithm”.

The fitness of a chromosome i ($F(i)$) is the average tardiness of the jobs (i.e. the overall performance criteria) of the simulated permutation (schedule).

Possible selections are the fitness-proportional selection (or roulette wheel selection), the tournament and a specific percentage which is chosen accidentally. With the fitness-proportional selection the probability $P(i)$ of

selecting the i -th chromosome is given by
$$\frac{F(i)}{\sum_{i \in P} F(i)}$$

with population P . In the tournament selection,

n (a parameter) chromosomes of the actual population are accidentally selected and the one with the best fitness comes in a new set P . This operation is repeated until a certain number of chromosomes are chosen. If m chromosomes have the best fitness, but just $m' < m$ ones can put in the set than the m' ones are chosen by accident. A chromosome can be selected and chosen more than once. Then, P is the new population.

There are two classes of genetic operations: crossover and mutation. Crossover mixes two selected chromosomes of the current population to two chromosomes. It is applied with probability PC , which is usually high ($> 0,6$ is often recommended). Mutation

changes the position of one or more orders (genes) in a single permutation (chromosome). It is applied with probability PM, which is usually small (often $0,01 < PM < 0,1$ is recommended).

Some of the crossover operators described in (Werner 2013) are implemented, namely: order crossover (OX), cycle crossover (CX) and order based crossover (OBX). Also the mutation operators described in (Werner 2013) are implemented, namely: shift-mutation (also named as neighborhood or insertion neighbourhood), pairwise interchange neighborhood (also named as swap neighbourhood), API neighbourhood, and inversion neighbourhood; note: (Werner 2013) contains some theoretical results about mutation operators.

Formation of the new population is done by elite strategy: The best chromosome or a specific percentage of the best chromosomes is in the next population and all others are chosen by accident.

As stopping criteria a maximal number of iterations – i.e. generating of populations –, maximal runtime is available or after a percentage of the number of orders in a scheduling problem.

4. EXPERIMENTAL CALIBRATION OF THE GENETIC ALGORITHM

Prior to applying the genetic algorithm on the real world application, the parameters are chosen. For this, the genetic algorithm applied on some small generated test problems.

In order to use problems, which are comparable to the real world problem, 4 stations are being considered. The routings are created from a set of so called basis routings which are stated in table 2. The total net processing times of these basis routings covers the same range as the ones in the real world application. A concrete routing is created from one basis routing R, as routing 3 for example. The processing time for a station S, as station 2 for example, is created by a normal distribution whose mean value is the processing of S in R, also 200 minutes in the example, and the deviation is either 20% of the mean value (small deviation) or 75% of the mean value (large deviation); of course negative processing times are excluded – in both cases. The pool of orders is too small to effectively generate a part type sequence by the generating algorithm. Instead, the part types are generated by a uniform distribution, and the following three basic scenarios for the order release are randomly generated: all orders are realised in one, two or three periods, so that in each scenario the difference of the number of orders in the periods is at most one – note, that sometimes orders of previous periods are still in the production system. The number of orders vary between 8, 16 and 24.

The due dates were generated by a fix flow factor, so that under scheduling with FIFO the percentage of late jobs is 30%, 50%, 70% or 85% – resulting in $3 \cdot 4 \cdot 3 = 36$ combinations. For each operation in the five basis routings (see table 2) 4 processing times are

generated. This leads to $5 \cdot 4 \cdot 4 = 80$ routings. So, in total $36 \cdot 80 = 2880$ experiments are generated.

By using the basis routings and a uniform distribution of the parts, station 3 is the bottleneck station, because the sum of all operation times at station 3 is greater than these sums at the other stations. Because of the way alternative processing times are being generated for the stations, the sequence of stations due to this criterion (the sum of all operation times at a station) can change. In the real application each station is a bottleneck in a significant portion of the periods over a large horizon, because the products are non-uniform distributed in the demand over a large horizon. In order to ensure a comparable situation, about 60000 are generated. From those 2880 are chosen, so that in around 15% of the experiments station 1 is a bottleneck station, in around 30% of the experiments station 2 is a bottleneck station, in around 25% of the experiments station 3 is a bottleneck station, and in around 30% of the experiments station 4 is a bottleneck station (and of course, the other conditions are still fulfilled).

Routing / part type	Station 1	Station 2	Station 3	Station 4
1	100	50	50	10
2	150	100	100	150
3	100	200	150	200
4	200	150	300	150
5	250	250	250	200
Sum of operation times	800	750	850	710

Table 2: Basis routings for the creation of scheduling problems in minutes

The parameters of the genetic algorithm are varied as follows:

- Population size: 5, 10, 20, 50, 100, 200, 400, 600, 800 and 1000.
- Selection: fitness-proportional selection, the tournament and accidentally chosen a specific percentage. The parameter n in the tournament is 20%, 40% and 60% of the number of orders in a scheduling problem.
- Crossover: OX, CX and OBX.
- Crossover probability: 0.0, 0.1, 0.2, 0.3, 0.4 and 0.5.
- Mutation: shift-mutation, pairwise interchange neighborhood, API neighbourhood and inversion neighbourhood.
- Mutation probability: 0.0, 0.005, 0.01, 0.015, 0.02 and 0.03.
- Formation (elite strategy): the best chromosome or a specific percentage of the best chromosomes is in the next population and all others are chosen by accident. The percentage is: 5%, 10%, 15%, 20%, 30% and 40% of the number of orders in a scheduling problem.
- Stopping criteria: after 30%, 40%, 50% and 60% of the number of orders in a scheduling problem.

Initial population is generated either by accident or by one or more priority rules. Over the last decades, many priority rules are suggested and analysed and due to the dynamic environments in industrial practise, priority rule are still analysed in many studies on scheduling – one example of a recent one is (El-Bouri 2012) – and they are often used in industrial practise. An investigation about priority rules to this real world application is presented by the author in (Herrmann 2013). This investigation shows that the calculation of the processing time of an order, which is assigned to the flow shop next, is critical for the performance of the priority rule. The processing time depends on the next jobs on the flow shop. In (Herrmann 2013) it is shown that a calibration of such a tail is possible and with this constant tail the priority rules delivers often better results. Especially, the most successful priority rules benefits from this setting.

The following priority rules – as in (Engell et al. 1994) or (El-Bouri 2012) – are used. In the definition t is the current time, f_i the due date of job i , tt_i the total processing time of i , calculated with tail or as sum of operation times (net processing time) and a low value is always preferred – without the named exceptions:

- First in first out = t_i , t_i is the arrival time in queue in front of the flow shop and last in first out = t_i , where a high value is preferred.
- Shortest processing time = tt_i .
- Longest processing time = tt_i ; here a high value is preferred.
- Earliest finishing time = $t_i + tt_i$.
- Earliest due date = f_i is here identical with earliest operational due date (because only the first station is scheduled) and modified earliest due date = $\max\{t + tt_i, f_i\}$.
- Slack = $f_i - t - tt_i$ is identical with slack per remaining number of operation = $\frac{f_i - t - tt_i}{M}$ and slack per remaining processing time = $\frac{f_i - t - tt_i}{tt_i}$.
- Truncated shortest processing time with parameter r = $\min\left\{tt_i + r, \frac{f_i - t - tt_i}{M}\right\}$.
- CR+SPT = $\begin{cases} \frac{f_i - t}{tt_i}, & f_i - t - tt_i > 0 \\ tt_i, & f_i - t - tt_i \leq 0 \end{cases}$.
- CR = $\frac{f_i - t}{tt_i}$.
- RR = $(f_i - t - tt_i) \cdot e^{-\eta} + e^{\eta} \cdot tt_i$, (by (Raghu and Rajendran 1993)) with utilisation level η of the entire flow shop defined by $\eta = \frac{b}{b+j}$ with b being

the busy time and j being the idle time of the entire flow shop.

- RM (by (Rachamadugu and Morton 1982)) with priority index: $\frac{1}{\pi_i} \cdot e^{-\frac{k}{i} \cdot \max\{f_i - t - tt_i, 0\}}$ with either local processing time costing $\pi_i^l = tt_i$ (called RM local) or global processing time costing $\pi_i^g = \sum_{i \in U_i} tt_i$, where U_i is the set of unfinished jobs in the pool of orders, excluding job i (called RM global).

Since mean tardiness is most important these small experiments are just evaluated by this criterion.

An analysis of the schedules determined by the genetic algorithm shows in some cases that an improvement is achieved by an empty station. Especially in a rolling scheduling there occur cycles of orders in which at least one order starts in the next period and another starts in the previous period and often it is beneficial to split this cycle in 2; so an empty station occurs. The procedure of priority rules cause in such cases an empty station and outperforms the genetic algorithm even if the sequences of orders calculated by a priority rule is in the initial population. Technically, an empty station in the genetic algorithm is achieved by an artificial order i whose duration time on each station is zero and whose release date is less than the release dates of all normal jobs, so i can start immediately. By a huge due date of i , no tardiness occur; thus, there is no effect on the objective function. Of course, such an artificial order could be not only beneficial at the end of a period, but also in between; even with high workloads (or time pressure, respectively), which occurs in the following experiments quite often. A preliminary study shows that 12 such artificial orders are sufficient for all experiments; also for the ones with the real word application.

The experiments shown that fitness-proportional selection outperforms tournament selection. In any case it is very beneficial that 20% of the best chromosomes survive by elite strategy. The population size has a large impact on the performance. An increase of the above named numbers until 400 causes a decrease, at the beginning a significant decrease, of the mean tardiness. A further increase (from 400) causes a (moderate) increase of the mean tardiness. An initial population of chromosomes generated by one or more priority rules is outperformed by a population of accidentally generated chromosomes. Such an initial population is improved by a mixture of both ways of determining an initial population. The best results are achieved by 15 copies of a chromosome generated by a single priority rule in the initial population. This is executed for each of the above mentioned priority rules. Calculating the processing time via a fix tail is better than using the net processing time. In addition, a sequence of orders is determined by the NEH heuristic with mean tardiness as

objective function and for presorting the list of orders each of the above named priority rules are used. Then, the initial population is filled up with accidentally generated chromosomes.

A major impact has the probability of crossover and mutation. As reported in other publications as well a high crossover probability is beneficial. The best mean tardiness is achieved by a probability of 0,9; it may be noted that the results by a probability of 0,4 is 4,4% above and by a probability of 0,4 it is 13,4% above. Compared to this the impact of the mutation probability is less important. The best value is achieved by a low probability of 0,005 which is just 0,74% better than the one by a probability of 0,02. These values are measured for the order crossover operator and the inversion neighbourhood mutation operator. This combination of operators (for crossover and mutation) delivers also for other probabilities (of these two operators) the best mean tardiness.

In any case a large number of iterations is beneficial, because with the elite strategy very good solutions survive over the generations. A stopping after 30% of the number of orders in a scheduling problem delivers nearly in the all cases the smallest mean tardiness.

4. COMPUTATIONAL RESULTS

The real world application is simulated for the sequence of orders explained in section 2. If an assignment of an operation on the first station ends in the next period t , the orders of period $(t+1)$ is realised, so that the orders of the periods t and $(t+1)$ are known. Average tardiness (T_{Mean}) and standard deviation of the tardiness (T_{σ}) reach a steady state by a simulation horizon of 5000 periods.

Due to the results published in (Herrmann 2013) genetic algorithm (GA) with the parameter setting due to the above calibration is compared to the results of the best priority rules, which are RR and RM local. The results shown in Table 3 are average objective values relative to the solutions of GA set to 100%; thus, e.g., the solutions generated by RR rule for time pressure 30% were about 51% above GA on the average.

Table 3: Relative performance measures of the best priority rules compared to the genetic algorithm (GA)

Rule	Time pressure			
	30%	50%	70%	85%
T_{Mean}				
GA	100	100	100	100
RR	1,51	1,41	1,33	1,21
RM local	1,75	1,69	1,51	1,4
RM global	2,93	2,51	2,2	1,99
T_{σ}				
GA	100	100	100	100
RR	1,21	1,19	1,13	1,09
RM local	2,4	1,9	1,63	1,51
RM global	3,5	2,1	1,64	1,42

The genetic algorithm outperforms the priority rules significantly. At the company site the benefit would even much higher, because their scheduling procedure is much simpler (than these priority rules).

Without the exception of the priority rules for generating chromosomes for the initial population, just standard operators are used in the genetic algorithm. A first analysis of optimal schedules of the above test problem shows that, for example, outliers in the cycle times are avoided, but priority rules have them often. It could be beneficial to have chromosomes with such a property in a (initial) population and operators who use them.

5. CONCLUSIONS

This paper presents a real world flow shop scheduling problem with more restrictive restrictions than the ones normally regarded in literature. Despite of standard operators in the genetic algorithm of this publication, it outperforms the improved priority rules in (Herrmann 2013) substantially. A further improvement seems to be possible by integrating specific properties of very good schedules, especially in the operators.

More technical restrictions in companies occur by limited resources, like the available number of coils or assembly ground plates, and workers for the manual tasks causes other schedules to be optimal or at least very good. Such requirements are also left to future investigations.

REFERENCES

- Januario, T.; J. Arroyo and M. Moreira. 2009. "Genetic Algorithm for Tardiness Minimization in Flowshop with Blocking". In: N. Krasnogor, B. Melián, J. Moreno, J. Moreno-Vega, D. Pelta (Editors): "Nature Inspired Cooperative Strategies for Optimization (NICSO 2008)". Studies in Computational Intelligence Volume 236, 2009, 153 – 164.
- El-Bouri, A. 2012. "A cooperative dispatching approach for minimizing mean tardiness in a dynamic flowshop". *Computers & Operations Research*, Volume 39, Issue 7 (July), 1305 – 1314.
- Chaudhry, I.; S. Mahmood. 2012. "No-wait Flowshop Scheduling Using Genetic Algorithm". In Proceedings of the World Congress on Engineering 2012 Vol III, WCE 2012, July 4 - 6, 2012, London, U.K..
- Engell, S.; F. Herrmann; and M. Moser. 1994. "Priority rules and predictive control algorithms for on-line scheduling of FMS". In *Computer Control of Flexible Manufacturing Systems*, S.B. Joshi and J.S. Smith (Eds.). Chapman & Hall, London, 75 – 107.
- Framinan, J.M.; R. Leisten; and C. Rajendran. 2003. "Different initial sequences for the heuristic of Nawaz, Ensore and Ham to minimize makespan, idletime or flowtime in the static permutation flowshop sequencing problem". *International Journal of Production Research*, 41, 121 – 148.

- Grabowski, J. and J. Pempera. 2007. "The permutation flow shop problem with blocking. A tabu search approach.". *Omega*, 35 (3), 302 – 311.
- Hall, N.G. and C. Sriskandarajah. 1996. "A survey of machine scheduling problems with blocking and no-wait in process". *Operations Research* 44 (3), 510–525.
- Herrmann, F. 2013. "Simulation based priority rules for scheduling of a flow shop with simultaneously loaded stations". In: Proceedings of the 27th EUROPEAN Conference on Modeling and Simulation, May 27th – 30th, 2013, Ålesund, Norway.
- Jacobs, F.R.; W. Berry; D. Whybark; T. Vollmann. 2010. "Manufacturing Planning and Control for Supply Chain Management". McGraw-Hill/Irwin (New York), 6 edition.
- Lawrence, S. and T. Morton. 1993. "Resource-constrained multi-project scheduling with tardy costs: Comparing myopic, bottleneck, and resource pricing heuristics.". *European Journal of Operational Research* 64, 168 – 187.
- McCormick, S.T.; M.L. Pinedo; S. Shenker; and B. Wolf. 1989. "Sequencing in an assembly line with blocking to minimize cycle time". *Operations Research*, 37 (6), 925 – 935.
- Nawaz, M.; E.E. Enscore; and I. Ham. 1983. "A heuristic algorithm for the m-machine, n-job flow sequencing problem". *Omega*, 11(1), 91 – 95.
- Pan, Q.; and L. Wang. 2012. "Effective heuristics for the blocking flowshop scheduling problem with makespan minimization". *Omega*, 40 (2), 218 – 229.
- Rachamadugu, R.M.V. 1987. "Technical Note –A Note on the Weighted Tardiness Problem". *Operations Research*, 35, 450 – 452.
- Rachamadugu, R.V. and T.E. Morton. 1982. "Myopic heuristics for the single machine weighted tardiness problem". Working Paper No. 28-81-82, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA.
- Raghu, T.S. and C. Rajendran. 1993. "An efficient dynamic dispatching rule for scheduling in a job shop". *International Journal of Production Economics* 32, 301 – 313.
- Rajendran, C and O. Holthaus. 1999. "A comparative study of dispatching rules in dynamic flowshops and job shops". *European Journal of Operational Research*, 116 (1), 156 – 170.
- Ribas, I.; R. Companys; and X. Tort-Martorell. 2011. "An iterated greedy algorithm for the flowshop scheduling problem with blocking". *Omega*, 39, 293 – 301.
- Ronconi, D.P. and V.A. Armentano. 2001. "Lower Bounding Schemes for Flowshops with Blocking In-Process". *Journal of the Operational Research Society*, 52 (11), 1289 – 1297.
- Ronconi, D.P. 2004. "A note on constructive heuristics for the flow-shop problem with blocking". *International Journal of Production Economics*, 39 – 48.
- Ronconi, D.P. 2005. "A branch-and-bound algorithm to minimize the makespan in a flowshop with blocking". *Annals of Operations Research*, (138), 53 – 65.
- Ronconi, D. and L. Henrique. 2009. "Some heuristic algorithms for total tardiness minimization in a flow shop with blocking". *Omega*, 37 (2), 272 – 281.
- Russel, R.S.; E.M. Dar-El; and B.W. Taylor. 1987. "A comparative analysis of the COVERT job sequencing rule using various shop performance measures". *International Journal of Production Research*, 25 (10), 1523 – 1540.
- Taillard, E. 1993. "Benchmarks for basic scheduling problems". *European Journal of Operational Research*, 64 (2), 278 – 285.
- Vepsalainen, A.P. and T.E. Morton. 1987. "Priority rules for job shops with weighted tardiness costs". *Management Science* 33/8, 95 – 103.
- Voß, S. and A. Witt. 2007. "Hybrid Flow Shop Scheduling as a Multi-Mode Multi-Project Scheduling Problem with Batching Requirements: A real-world application.". *International Journal of Production Economics* 105, 445 – 458.
- Wang, X. and L. Tang. 2012. "A discrete particle swarm optimization algorithm with self-adaptive diversity control for the permutation flow shop problem with blocking". *Applied Soft Computing*, (12, 2), 652 – 662.
- Wang, L.; Q.-K. Pan; and M.F. Tasgetiren. 2011. "A hybrid harmony search algorithm for the blocking permutation flow shop scheduling problem". *Computers & Industrial Engineering*, 61 (1), 76 – 83.
- Wang, C.; S. Song, S.; J.N.D. Gupta; and C. Wu. 2012. "A three-phase algorithm for flowshop scheduling with blocking to minimize makespan". *Computers & Operations Research*, 39 (11), 2880 – 2887.
- Werner, F. 2013. "A survey of genetic algorithms for shop scheduling problems". In: Heuristics. - New York, Nova Publishers, 133 – 160.

AUTHOR BIOGRAPHY



Frank Herrmann was born in Münster, Germany and went to the RWTH Aachen, where he studied computer science and obtained his degree in 1989. During his time with the Fraunhofer Institute IITB in Karlsruhe he obtained his PhD in 1996 about scheduling problems.

From 1996 until 2003 he worked for SAP AG on various jobs, at the last as director. In 2003 he became Professor for Production Logistics at the University of Applied Sciences in Regensburg. His research topics are planning algorithms and simulation for operative production planning and control. His e-mail address is Frank.Herrmann@OTH-Regensburg.de and his Web-page can be found at www.hs-regensburg.de/frank.herrmann

MapReduce Based Experimental Frame for Parallel and Distributed Simulation Using Hadoop Platform

Byeong Soo Kim, Sun Ju Lee, and Tag Gon Kim
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, 305-701, Republic of Korea
E-mail: {bskim, sjlee}@smlab.kaist.ac.kr
tkim@ee.kaist.ac.kr

Hae Sang Song
Department of Computer Engineering
Seowon University
Cheongju, 361-742, Republic of Korea
E-mail: hssong@seowon.ac.kr

KEYWORDS

MapReduce, experimental frame, parallel and distributed simulation, Hadoop, system analysis.

ABSTRACT

Simulation-based experiment of complex systems is a time consuming-job. Parallel and distributed simulation is one of the methods to reduce the simulation time. To simulate and analyze the system with this method, it is required to design a suitable experimental frame. Therefore, this paper proposes a MapReduce based experimental frame for the parallel and distributed simulation. Because Hadoop MapReduce is the most widely used parallel and distributed computing platform, we use it to design the experimental frame. In our work, the 'map' of MapReduce automatically generates and simulates the system, and the 'reduce' of MapReduce collects and analyzes the result. We can reuse the existing large scale Hadoop clusters without any modification of the platform, so it is easy to set-up and use the experimental frame. This paper presents an air defense simulation to show the usage and speed up with a 16-node Hadoop cluster.

INTRODUCTION

To analyze characteristics of a system, relationships must be drawn between input parameters and performance indices of the system. The more complex the system is, the more researchers need time and effort to draw how the inputs affect the performance indices. For example, assume that there are various parameters like number of missiles, speed of missile, accuracy of missile, range of radar, attack power, and decision making time for the simulation of an air defense system (Cho et al. 2007). If each parameter has five levels, the system could have over 10,000 scenarios with full factorial design (Antony 2003). Then, there needs to be over 10,000 minutes to simulate all the scenarios of the system, even if each simulation takes only one minute. It is also a time-consuming job to collect and analyze the results after the simulation.

Therefore, many researches have attempted faster simulation methods to remedy this problem. Generally, parallel and distributed simulation approaches have been widely used to reduce time-consuming phenomenon (McGeoch 1992). To simulate the system with parallel

and distributed environment, it is required to design an appropriate experimental frame. An experimental frame is a specification of the conditions under which the system is experimented with (Zeigler et al. 2000). It is composed of a generator, which generates the input segments and a transducer, which collects and analyzes the output segments of the system.

In simulation fields, there is research adapting parallel and distributed simulation techniques for faster data collection of the simulation: DEXSim (Choi et al. 2014), CR-PADS (Bononi et al. 2005), and mJADES (Rak et al. 2009). They provide efficient simulation environments with best use of distributed hardware resources, however they did not consider faster data analysis. In other words, they did not provide an experimental frame for faster data collection and faster data analysis. Furthermore, since the previous studies were developed in their specific environment, they spend much time and cost to expand the distributed machines.

In our approach, parallel and distributed simulation is used to design and simulate a system; it is also used to gather and analyze the results after the simulation. We use the Hadoop platform for implementation of the proposed work for the convenience of environment construction. Hadoop is the most widely used platform for distributed computing (White 2012). It is a scalable, common, and reliable platform compared to other platforms used for the previous studies. Furthermore, MapReduce, a distributed computing framework of Hadoop, is appropriate to adapt the concept of generator and transducer. Although there is research about simulation using MapReduce (Decraene et al. 2011; Jakovits et al. 2011; Pratz and Xing 2011), there is no research about experimental frames for the simulation of legacy simulators.

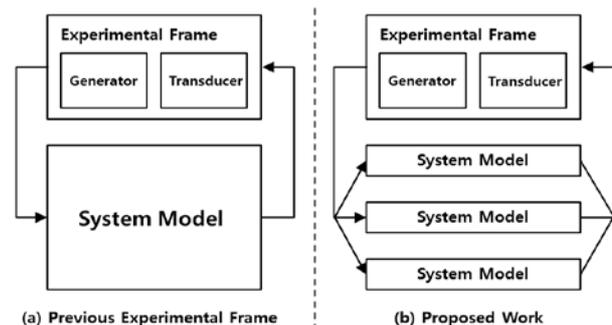


Figure 1: Previous and Proposed Experimental Frame

In this paper, we propose a new experimental frame for parallel and distributed simulation using the Hadoop platform (Figure 1-(b)). The proposed work provides an experimental frame for efficient experimental design and result analysis. It can also be used for simulation-based optimization (Hong et al. 2013). Because it reuses the existing large-scale Hadoop clusters without any modification to the platform, users do not need to set the distributed environment. Basically, Hadoop provides reliability and powerful load balancing; users have only to take advantage of Hadoop platform.

This paper is organized as follows. Background materials about the Hadoop platform and Hadoop Streaming are briefly introduced. Then our proposed experimental frame using the Hadoop platform and simulation process are described. Finally, a case study using an air defense simulator is provided.

HADOOP

Hadoop is a representative big data platform developed by the Apache Software Foundation. It is an open source implementation for reliable, scalable, large scale distributed computing (White 2012). Hadoop consists of MapReduce and Hadoop Distributed File System (HDFS). MapReduce is a distributed computing framework for large scale data processing. HDFS is a distributed file system that stores data reliably using commodity machines (Shvachko et al. 2010). The Hadoop platform is fault-tolerant for hardware and network failures, and provides efficient resource management.

MapReduce

MapReduce is a framework for large-scale distributed data processing based on the divide and conquer paradigm. MapReduce works by breaking the processing into map and reduce (Dean and Ghemawat 2008). Map and reduce are executed in parallel on the different machines within the Hadoop cluster by MapReduce framework. Map performs filtering and sorting operations, and reduce performs summary operations. The user can specify map/reduce functions, and types of input/output.

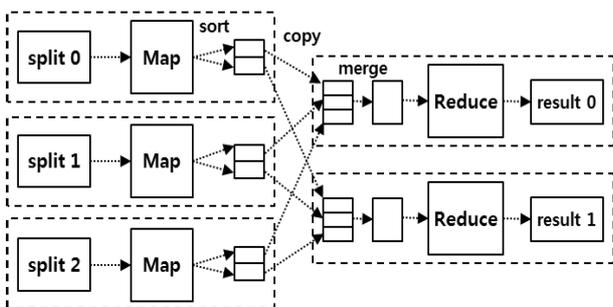


Figure 2: Overall Process of MapReduce

Figure 2 illustrates the overall process of MapReduce. Input data stored on the HDFS are split into fixed-size

blocks, and each block is allocated to a map. Then user-specified map processes each key-value pair in the block; and outputs the result as a list of key-value pairs. The output of the map is partitioned by the key, and the grouped records are transferred to the different reduces, respectively (called shuffle). Then, the transferred records are merged and sorted in the node which a reduce task located. Each reduce sequentially reads key-value pairs, and processes them by the user-specified reduce function. Finally, the output records of the reduce are written to the HDFS.

$$\begin{aligned} \text{Map } (k_1, v_1) &\rightarrow \text{list } (k_2, v_2) \\ \text{Reduce } (k_2, \text{list } (v_2)) &\rightarrow \text{list } (v_3) \end{aligned}$$

Hadoop Streaming

Because MapReduce applications are executed in the form of source code, it is difficult to run an executable legacy simulator on the MapReduce framework. Therefore, an interface is required to run one on the MapReduce framework. Developers can implement the interface directly, but for convenience, the Hadoop platform provides a utility called Hadoop Streaming.

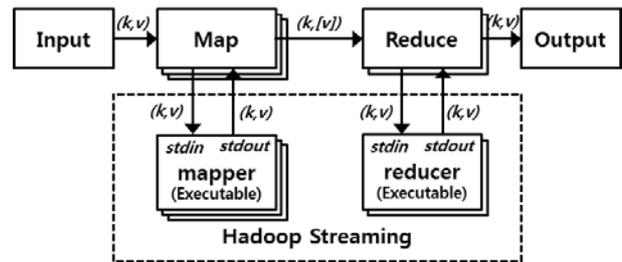


Figure 3: Overview of Hadoop Streaming

Hadoop Streaming is a Hadoop utility (application) which allows creating and running MapReduce jobs with any executable program as the Mapper and Reducer. Executable programs communicate with Hadoop Streaming through Unix streams (Figure 3). They read the input key-value pairs from standard input (stdin) line by line, and emit the output key-value pairs to standard output (stdout). In this paper, we use Hadoop Streaming to implement the proposed work.

MAPREDUCE BASED EXPERIMENTAL FRAME

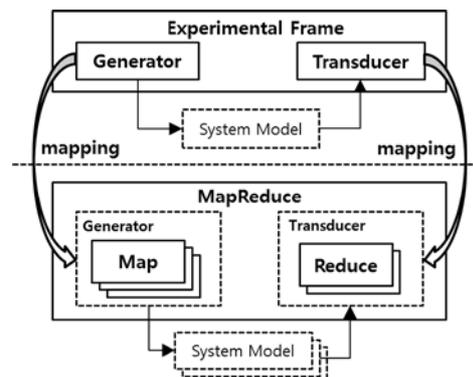


Figure 4: Conceptual Mapping to MapReduce

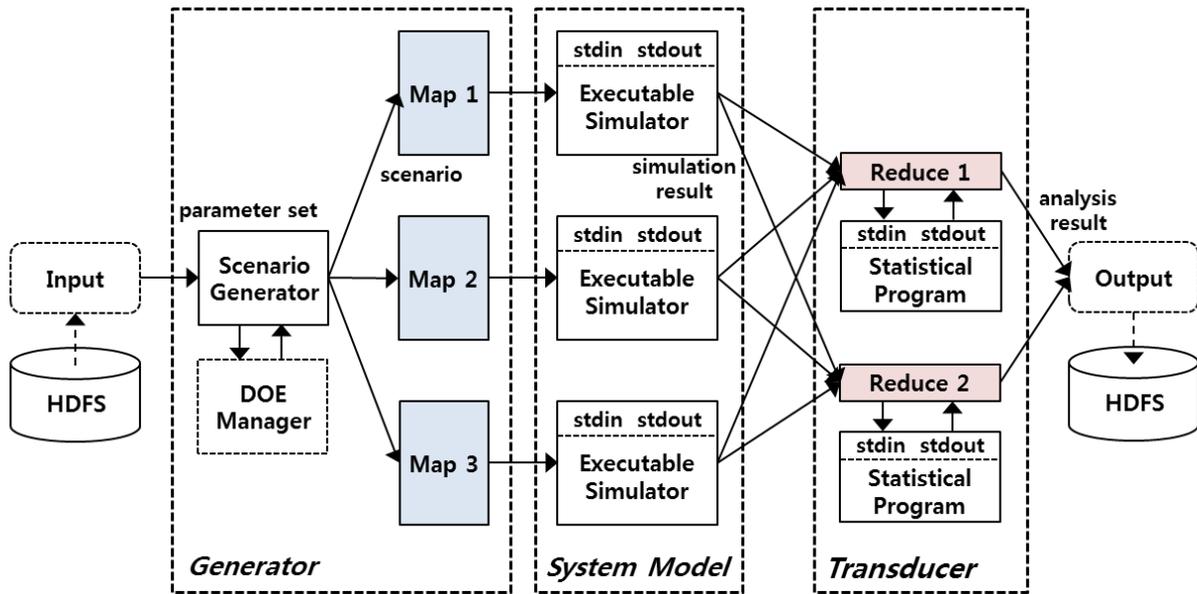


Figure 5: Overall Structure of Experimental Frame

In this section, we propose an implementation of an experimental frame adapting Hadoop MapReduce. Figure 4 shows the conceptual mapping between the experimental frame and the MapReduce framework. In our approach, map performs the role of generator which generates scenarios and allocates them to the system model. The reducer performs the role of transducer which collects the output of simulations and analyzes the results.

Overall Structure

The generator of the proposed experimental frame is composed of scenario generator, design of experiment (DOE) manager, and map of MapReduce framework (Figure 5). The scenario generator makes scenarios from input parameter set using DOE manager. The DOE manager handles the design of experiment, but in this paper, only full-factorial design (Antony 2003) can be used. Later, it is possible to apply several design of experiment methods extending the DOE manager. The outputs of the scenario generator (all scenarios) are automatically split into an individual scenario by MapReduce framework, then each scenario is allocated to each map. MapReduce framework assigns a map to each CPU core of individual machine in Hadoop, and provides efficient load balancing and resource management. Consequently, efficient and faster data collection are possible with parallel and distributed environment of Hadoop.

The transducer is composed of statistical program and reduce of MapReduce framework. The statistical program, which can be a commercial (e.g., R project for statistical computing) or a user-developed application, processes and analyzes the simulation output. The results of the analysis can be statistical values of the simulation, optimized input parameters or extracted internal state. The reduce collects and saves the output

to the HDFS. It is executed in parallel like the map, therefore faster data analysis is possible using the experimental frame. In this proposed work, the legacy simulator has to receive an input scenario from stdin, and emit the output to stdout in the form of key-value pair. It is the same with the statistical program.

Simulation Process

The simulation process is composed of two phases: the preparatory phase, and the main phase (Figure 6). The preparatory phase generates scenarios and sets up the experimental environment. The main phase executes the simulator and analyzes the simulation result. The detail processes are as follows.

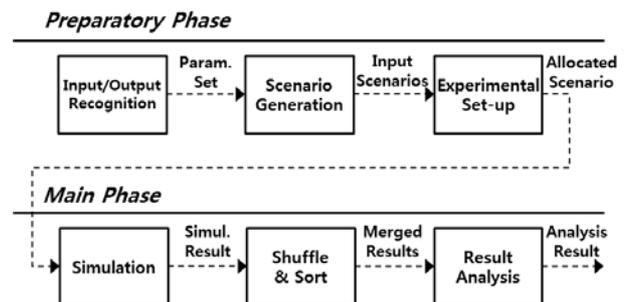


Figure 6: Simulation Process: Preparatory Phase and Main Phase

Preparatory Phase

In the preparatory phase, we specify object/attribute and performance index following the objective of the simulation. We use the Object-Performance Index (OPI) matrix to recognize the relation between the object and the performance index (Kim and Sung 2007) and make the Parameter Set shown in Table 1. The experimental frame automatically generates scenarios adapting the design of the experiment and writes the scenario file to

the HDFS. It also saves the simulation model and simulation engine (in this paper, DEVS simulation model and DEVSimLinux binary files) to the local file system (LFS) of each node. Then, the experimental frame sets up the experimental environment: It sets the total number of map tasks following the number of scenarios, the number of reducers, and the number of task slots. Finally, the MapReduce framework splits the scenario file into the number of maps and allocates each scenario to each map task.

Table 1: Specification of Input / Output

Type	Example
Parameter Set	(Attribute 1 : range of value, Attribute 2 : range of value, ... Performance Index : threshold value)
Input Scenario	(Scenario 1, Parameter 1 : value, Parameter 2 : value, ...)
Simulation Result	(Scenario 1, Result 1 : value, Result 2 : value, ...)
Analysis Result	(Scenario 1, Parameter 1: value, ... Performance Index : value)

Main Phase

In the main phase, each map executes the simulator stored in the LFS with the allocated scenario. The simulator reads the scenario in the form of the key-value pair. After the simulation is finished, it emits the result of the simulation in the form of the key-value pair to its own map. Then the simulation results of all the maps are sent to the reduce through the shuffle process of MapReduce. The reduce merges and sorts the results from the maps, and the developed statistical program analyzes the results. Finally, they are merged and written in the HDFS by the MapReduce framework. The main phase is automatically performed by the MapReduce framework, so the user does not need to manage the simulation after the execution of the experimental frame.

CASE STUDY: AIR DEFENSE SIMULATOR

This section presents a case study to show the efficiency of the proposed experimental frame. The experiment was conducted on a homogenous Hadoop cluster of 16 machines, which consisted of one master machine and 15 slave machines. Each machine had quad-core Intel i5-3550 CPUs running at 3.3 GHz, Samsung DDR3 4 GB RAM, and Samsung HDD 500 GB, running on Ubuntu-12.04 32bit. We used Hadoop-1.1.2. The machines were connected to a gigabit Ethernet switch with two trunked gigabit ports per machine.

Target Simulator

The target simulator is an air defense system simulator which detects and nullifies missiles from an enemy. It is modeled using DEVS formalism, and is running on the DEVSimLinux, discrete event system simulation engine

(Kim et al. 2011). The simulator is composed of four parts; missile, radar, weapon, and C2A (Command & Control and Alert) system model. The C2A system receives target information from radars, makes decisions based on algorithms, and sends attack orders to the air defense weapon systems.

The simulator represents the situation of defending a base against missiles and analyzes the effectiveness of the air defense system for various parameters. When missiles are fired, installed radars detect the missiles and send the information to the C2A system. Then the C2A system assigns weapon systems according to the algorithms and orders an attack to defend the base. Finally the simulator measures the defense rate in accordance with the success or failure of the attack.

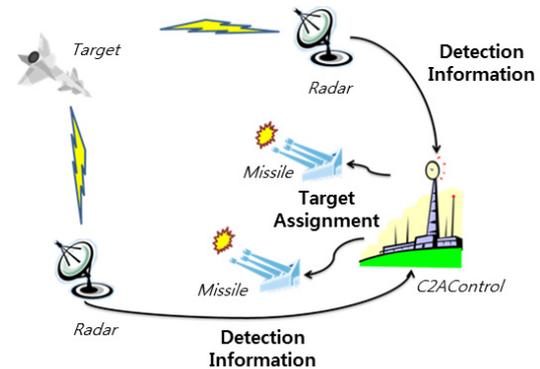


Figure 7: Air Defense Simulator

Experimental Design

We designed experiments to acquire the scenarios whose defense success rate was more than 80% for the various parameters. To find the desired scenarios, full-factorial design is needed for the input parameters and its values. So, the generator was implemented to do full-factorial design for the input. Each scenario is executed 30 times to calculate its defense success rate. Table 2 shows six parameters and four values per parameter, as an input for the generator. Therefore, the generator makes a total of 4,096 scenarios and 122,880 simulation runs are required.

Next, we designed a transducer to find the desired scenarios among all scenarios. The transducer can utilize statistical programs, but in this case study, we implemented just a simple filter to find the desired scenarios.

Table 2: Parameter Set of Air Defense Simulator

Parameter	Value	Level
Radar Detection Range (km)	3 ~ 6	4
Number of Radars	1 ~ 4	4
Period of C2A (sec)	1.0 ~ 2.5	4
Weapon Range (km)	1.5 ~ 3.0	4
Weapon Accuracy Rate (%)	60 ~ 90	4
Number of Weapons	3 ~ 6	4
Total Scenarios	4,096 (=4 ⁶)	

Experimental Result

File: [/output/streaming/part-00000](#)

Goto:

[Go back to dir listing](#)
[Advanced view/download options](#)

```
No scenario is assigned to this task!!
Scenario1 : 0XX000X000X000X000000X0X00000 (AverageTime: 85.34, SurvivalRatio: 0.7333)
Scenario2 : XXX000X000X000X000000X00X0X0X0 (AverageTime: 85.4733, SurvivalRatio: 0.6)
Scenario3 : 00X000000000000000000000000000 (AverageTime: 85.2867, SurvivalRatio: 0.833)
Scenario4 : 0X00000000000000000000000000000 (AverageTime: 85.4533, SurvivalRatio: 0.666)
Scenario5 : 0000000000X000X000000000000000000 (AverageTime: 85.28, SurvivalRatio: 0.8333)
Scenario6 : 0000000000000000000000000000000 (AverageTime: 85.3267, SurvivalRatio: 0.8)
Scenario7 : 000000X0X0000000000000000000000 (AverageTime: 85.3067, SurvivalRatio: 0.833)
Scenario8 : 0000X00X000X00X0000000000000000 (AverageTime: 85.4, SurvivalRatio: 0.6)
```

Figure 8: Experimental Result of Air Defense Simulator Stored in HDFS

After the experiment completed, we got the result stored in HDFS as shown in Figure 8. Since Hadoop provides a monitoring tool for MapReduce and HDFS, it is convenient to check the progress of the experiment using this tool. The center of Figure 8 shows the experimental results: scenario number, performance index and statistical values such as simulation time. We can easily find the desired scenarios from the large number of scenarios.

Table 3: Execution Time of Total Experiment

	Hadoop	Single Node
Number of Scenarios	4096	4096
Execution Time (sec)	753	$5 \times 4096 = 20480$
Execution Time with 30 Cores (sec)	753	$20480 \div 30 = 680.67$
Speed up (times)	$20480 \div 753 = 27.20$	
Utilization Rate	$680.67 \div 753 = 0.90$	

Also, we can analyze how much the proposed experimental frame can reduce the execution time. We compared total execution time of the Hadoop platform with single node. In this experiment, two simulators can be executed simultaneously in one node because the number of map slots is 2. So, theoretical speed up of the proposed work should be 30 times with 15 slave nodes. However, the maximum speed up was only about 27 times as shown in Table 3 due to overhead of the Hadoop platform. Also, Figure 9 shows that the more the scenarios increase, the more the utilization of the experimental frame also increases. This is because the ratio of Hadoop overhead becomes smaller, as the number of scenarios increases. Consequently, we know that the proposed experimental frame is more efficient for the simulation with larger numbers of scenarios.

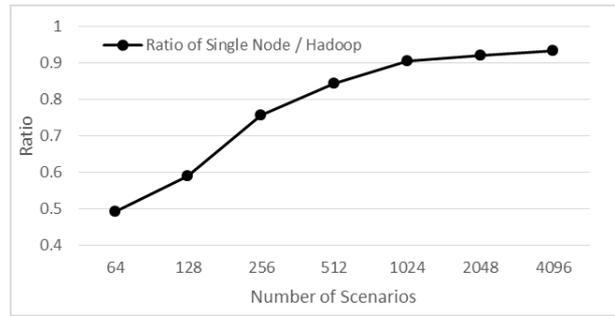


Figure 9: Utilization Rate according to Number of Scenarios

CONCLUSIONS

This paper proposes a MapReduce based experimental frame for parallel and distributed simulation using the Hadoop platform. Because simulation with a large number of scenarios requires much time to execute and analyze, the proposed work provides a generator for scenario generation and distributed execution of the simulator, and a transducer for distributed result analysis. In the proposed experimental frame, each generator/transducer is assigned to a CPU core by the MapReduce framework; faster data collection and data analysis are possible. The proposed work is also compatible with any modification or set-up of existing Hadoop cluster.

In this paper, we apply an air defense simulator which was developed in DEVS formalism to show the usefulness of the proposed work. A total of 4,096 scenarios were simulated on the 16-node Hadoop cluster with 30 map slots. The proposed experimental frame is only 27-times faster than a single node due to overhead of the Hadoop platform.

For further work, we will implement the DOE manager supporting several designs of experiment methods, and the statistical program analyzing these design of experiment methods. Also, we will research optimization methods using our proposed work.

ACKNOWLEDGEMENT

This work was partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract. (UD110006MD)

REFERENCES

- Antony, J. 2003. *Design of Experiments for Engineers and Scientists*, Butterworth-Heinemann.
- B.G. Cho; D.Y. Kim; S.H. Kim; C. Youn. 2007, "Real-Time Distributed Simulation Environment for Air Defense System Using A Software Framework". *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, Vol. 4 No. 2 (April), 64-79.
- Bononi, L.; M. Bracuto; G. D'Angelo; and L. Donatiello. 2005. "Concurrent replication of parallel and distributed simulations". In *Proceedings of Workshop on Principles*

- of *Advanced and Distributed simulation* (Jun.1-3), 234-243.
- Choi, C.B.; K.M. Seo; and T.G. Kim. 2014, "DEXSim: an experimental environment for distributed execution of replicated simulators using a concept of single simulation multiple scenarios." *SIMULATION: Transaction of The Society for Modeling and Simulation International*, Accepted to be published, Jan.
- Dean, J., and S. Ghemawat. 2008. "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM*, Vol. 51 Issue 1 (Jan), 107-113.
- Decraene, J.; F. Zeng; M.Y.H. Low; W. Cai; Y.Y. Cheng; and C.S. Choo. 2011. "Evolutionary Design of Experiments using the MapReduce Framework". *SCSC '11 Proceeding of the 2011 Summer Computer Simulation Conference*, 76-83.
- Jakovits, P.; I. Kromonov; S.N. Srirama. 2011. "Monte Carlo linear system solver using MapReduce". *2011 Fourth IEEE International Conference on Utility and Cloud Computing* (Victoria, NSW, Dec.5-9), 293-299.
- Jeonghee Hong; Kyung-Min Seo; and Tag Gon Kim. 2013. "Simulation-based optimization for design parameter exploration in hybrid system: a defense system example," *SIMULATION: Transactions of The Society for Modeling and Simulation International*, Vol. 89, No. 3 (Jan), 362-380.
- Kim, T.G. and C.H. Sung. 2007 "Objective-driven DEVS Modeling Using OPI Matrix for performance Evaluation of Discrete Event Systems". In *Proceedings of the 2007 Summer Computer Simulation* (San Diego, USA, Aug.), 305-311.
- Pratx, G. and L. Xing. 2011 "Monte Carlo simulation of photon migration in a cloud computing environment with MapReduce." *Journal of Biomedical Optics*, Vol. 16(12) (Dec).
- Rak, M.; A. Cuomo; and U. Villano. 2012. "mJADES: Concurrent Simulation in the Cloud". In *Proceedings of 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems* (Palermo, Italy, Jul.4-6), 853-860.
- Shvachko, K.; H. Kuang; S. Radia; and R. Chansler. 2010. "The Hadoop Distributed File System". *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies* (Incline Village, NV, May 3-7), 1-10.
- Tag Gon Kim; Chang Ho Sung; Su-Youn Hong; Jeong Hee Hong; Chang Beom Choi; Jeong Hoon Kim; Kyung Min Seo; and Jang Won Bae. 2011 "DEVSim++ Toolset for Defense Modeling and Simulation and Interoperation," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, Vol. 8, No. 3 (July), 129 - 142.
- White, T. 2012. *Hadoop – The Definitive Guide*, O'REILLY®, YAHOO!® PRESS.
- Zeigler, B.P.; H. Praehofer; and T.G. Kim. 2000. *Theory of Modeling and Simulation*, Second Edition, Academic Press.

AUTHOR BIOGRAPHIES

BYEONG SOO KIM is a Ph D. candidate at the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST). His research interests include methodology for M&S of discrete event systems (DEVS), distributed simulation, and system design.

SUN JU LEE is a master's degree candidate at the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST). His research interests include methodology for M&S of discrete event systems (DEVS) and distributed file system.

TAG GON KIM is a professor at the Department of Electrical Engineering Korea Advanced Institute of Science and Technology (KAIST). He was the Editor-In-Chief for *Simulation: Transactions for Society for Computer Modeling and Simulation International* (SCS). He is a co-author of the text book, *Theory of Modeling and Simulation*, Academic Press, 2000. He published about 200 papers in M&S theory and practice in international journals and conference proceedings. He is very active in defense modeling and simulation in Korea.

HAE SANG SONG studied his MS.D and Ph.D courses in Electrical Engineering in Korea Advanced Institute of Science and Technology (KAIST). He worked for a couple of years in an R&D lab, IAE (Instituted of Advanced Engineering) in 1999-2000. He also worked in a venture company for about two years, and has been a professor of Dept. Computer Engineering of Seowon University, Korea, since 2002. His major interest resides in modeling simulation, analysis, and control of discrete event dynamic systems.

DERIVING A MATHEMATICAL MODEL OF A PAINT SHOP FROM DATA ANALYSIS

Thomas Husslein, Christian Danner, Markus Seidl,
Jörg Breidbach
Department Algorithm Design and Optimization
OptWare GmbH
Prüfeninger Str. 20, D-93049 Regensburg, Germany
E-Mail: thomas.husslein@optware.de

Wolfgang Lauf
University of Applied Sciences OTH Regensburg
Universitätsstr. 31, D-93053 Regensburg, Germany
E-Mail: wolfgang.lauf@oth-regensburg.de

KEYWORDS

Paint shop, stochastic modelling, data analysis

ABSTRACT

The authors deal with the optimization of production planning in the mixed-model assembly production. Modern paint shops are highly complex facilities with a multitude of interdependent process steps. In order to describe the occurring throughput and processing times, the behavior of these times is of great interest for production planning, a modeling of the paint shop is necessary. The disruptions appearing here, which are a major source for delays and reorganizations within the car plant, are due to the complex structure and various rework. In this work, an alternate formulation of a long term reliable model is discussed: The description of the paint shop based on a stochastic model which is directly derived from data analysis of the production data.

COMPLEX STRUCTURE OF A PAINT SHOP

Due to the many successive steps the production process of painting a body can be viewed as a multi stepped flow production (Spieckermann, 2002): Figure 1 gives a schematic overview over the major technological steps of a paint shop. In general this is reduced to the following steps:

Pre-treatment: Arriving from the body shop, the body gets degreased and all metal fragments that are remnants of the production process in the body shop get cleaned away. Also some corrosion protective substances are applied.

Base coat: At the base coating, the body is immersed in a bath of electrostatic particles that coat the body due to electrostatic forces. This is predominantly for protection from corrosion and for optimizing the color application. At the end of this step the body is heated

in an oven to permanently fix the coating particles on the body.

Underbody seam sealing: In this process step overlapping of the metal sheeting and gaps are sealed to prevent water intrusion. Additionally the underbody gets an extra coating against stone impact.

Filler coating: After the surface of the body has been cleaned of dust again, the filler coating is applied. The brightness of the filler coating is adjusted concerning the brightness of the final color. This is an additional protection against stone chip.

Top coating: The top coat is the layer of color that is determining the final color of the product.

Clear finish: After the top-coating the whole body is painted with a clear paint. Clear finish is used in order to protect the color against scratches and other environmental influences.

Final inspection and rework: In these steps the color is investigated for quality problems. In case of quality failures the painted body is transferred to the reworking, otherwise the painted body is clear and subsequently transferred to a storage unit. The reworking unit consists of several steps. In the inspection the severity of the problem and the specialized workstation for the repair is assessed. There are several work stations depending on the severity of the problem. The three major stations are paint removal, spot repair and extended completion. After a re-assessment of the quality of the repair work the bodies are re-introduces into the line at the appropriate position.

The detailed description of each of these production steps is very difficult. Parallel process steps and the building of uniquely-colored batches have to be described. Furthermore, the paint shop of a car manufacturer is a place of continuing change. So the

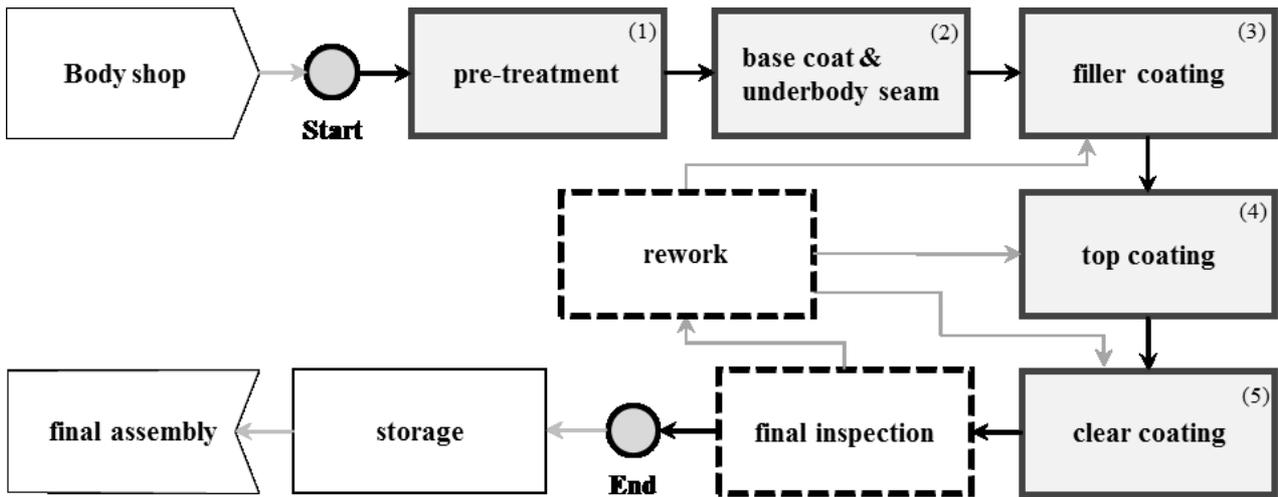


Figure 1: Basic Layout of a paint shop

formulation of long-term reliable model is nearly an impossible task.

PROBABILISTIC NATURE OF PAINT SHOP PROCESSES

Modern car production is organized by the pearl chain principle (Meyr, 2004; Weyer, 2002): The individual orders are lined up like pearls on a chain fixed in their sequence position relative to each other.

Complex processes in the final assembly are based on the production order planned. For the delivery processes just-in-time (JIT) and just-in-sequence (JIS) the stability of the production order is of central importance. As a consequence the description of the behavior of a paint shop is essential for optimal production planning and for using production control systems.

Parallelization of lines, processing of bodies in batches and rework due to system failures or violation of quality parameters cause turbulences of the pre-planned order. Especially, painted bodies which have to be reworked are separated from the assembly line and are reworked on specialized work stations.

A simulation model describing this process has to take into account this stochastic throughput.

From an abstract point of view, the paint shop can be regarded as a so called “black-box”. Bodies in white are entering the paint shop and colored bodies are leaving it. Each body has an individual throughput time depending on its path through the paint shop. The individual throughput times of the bodies can be easily obtained from data analysis.

DISTRIBUTION OF THROUGHPUT TIMES

The throughput time of a body i is defined as the time it takes to successfully pass through a workstation or a group of workstations (Arnold and Furmans, 2009). In this case we are interested in the throughput time through the paint shop which is defined as the timestamp a body leaves the paint shop minus the timestamp the body has entered the paint shop. The manufacturing information system of a car manufacturer stores for each produced car many production timestamps. By using the timestamps from the input and the output of the paint shop a top-down analysis is applied.

Through extensive data analysis, working time models and production interruptions have to be removed from the data set. By counting the bodies with defined throughput times random variables and corresponding distributions can be used to describe the throughput time. But, what is the correct distribution function?

To answer this question several distribution functions where fitted to the distributions from data analysis. To obtain the parameters of the distributions, on the one hand maximum likelihood estimators and on the other hand moments estimators have been used.

In general, empirical distributions of processing times are often skewed to the right, which is a result of the already mentioned production delays. So we have to concentrate on nonsymmetrical density functions.

A commonly used distribution is the exponential distribution.

$$F(x) = \begin{cases} \int_0^x 1 - e^{-\lambda y} dy & , x > 0 \\ 0 & , x \leq 0 \end{cases} \quad (1)$$

with a parameter $\lambda \in \mathbb{R}_{>0}$. The exponential distribution is practically only applied when standard deviation and mean are of roughly the same size. In order to compensate for this the distributions are generalized introducing new parameters. This leads to families of new distributions for example gamma distributions (Curry and Feldman, 2011. Manitz, 2005). The probability density function of the gamma distribution $GAM(\alpha, \lambda)$ is:

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & , x > 0 \\ 0 & , x \leq 0 \end{cases} \quad (2)$$

where Γ is the gamma function and the parameters α, λ are positive real numbers.

To study the effects of turbulence on the quality of the sequence and to take into account in the model later, it is necessary to calculate the differences from the target sequence. In order to measure the sequence variations between two stations, the relative order of deviation is calculated. This is the position number at the subsequent station minus the position at the previous station.

A typical distribution for the description of sequence variations is the lognormal distribution (Meißner, 2009). The lognormal distribution is defined by

$$F(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{1}{y} e^{-\frac{1}{2}\left(\frac{\ln(y)-\mu}{\sigma}\right)^2} dy & , x \geq 0 \\ 0 & , x \leq 0 \end{cases} \quad (3)$$

with parameters $\mu, \sigma \in \mathbb{R}, \sigma > 0$.

Also the normal distribution has to be mentioned. Although it is unsuitable for production processes that involve quality control and rework because of the symmetry, it is nevertheless an important distribution for production processes (Bayer et al., 2003). The normal distribution is defined by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \quad (4)$$

with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}, \sigma > 0$.

Another important issue is the mapping between sequence variation and throughput time. Using cycle time ZZ , average stock WIP and sequence variation RFA , it is possible to convert these values into each other (Meißner, 2009). For the throughput time M , one obtains:

$$M = (RFA + WIP) ZZ \quad (5)$$

In Figure 2, the densities of the measured and calculated cycle time are shown; they are almost identical (Danner, 2013).

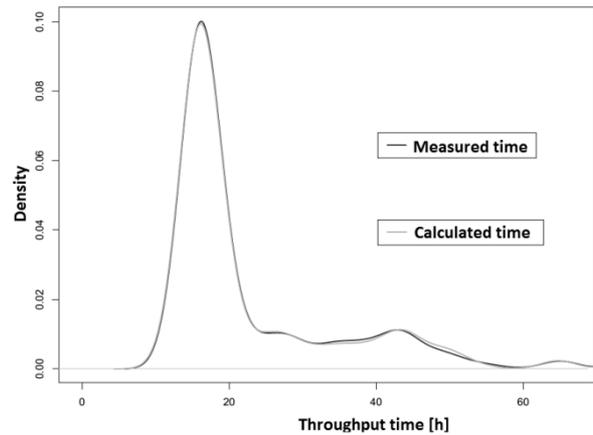


Figure 2: Calculated and measured throughput time

MODELLING THROUGHPUT TIME

In addition to understand and also to refine the distributions derived in the top-down analysis, we did a bottom-up analysis of the paint shop, based on detailed invariant points within the plant and the associated data.

By using the timestamps recorded by the manufacturing information system within the paint shop at each workstation, the paths of individual bodies through the paint shop are evaluated.

Using directed graphs an “empirical topological structure” of the paint shop can be visualized. An example is shown in Figure 3.

The analysis shows that the individual paths through the paint shop are branched and crossing each other in a complex fashion. Yet there are several nodes that all vehicles are passing through. The graph of the paint shop is divided by these common nodes into different sectors. The throughput time distribution is then examined in more detail for these sectors individually.

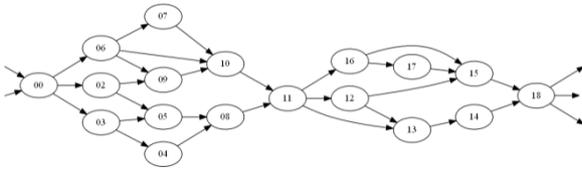


Figure 3: A part of paint shop structure

Data analysis of the throughput time distribution between the nodes suggests a possible separation of the paint process into five major sub-processes. These can be mapped to the big process steps initially described: Pre-treatment, Base coat, Filler coating, Top coating and clear finish, Final inspection and rework.

Throughput time has been analyzed and statistically clustered concerning the color of the top coating (Danner, 2013). For the throughput time we use $M(c)$ depending on the color c of the top coat. Typically, the number of colors is 20.

The throughput time for the paint shop is then leading to a refined model

$$M(c) = m(c) + \sum_{j=1}^5 X_j(c) \quad (6)$$

where $m(c)$ is the fixed time for the overall transport on conveyors in the paint shop and $X_j(c)$ is the random variable for the duration of production step j of color c (Danner, 2013). In order to incorporate the significant influence of rework on the distribution of the throughput time we split the process time $X_j(c)$ into a process time $X_{j1}(c)$ for bodies with no reworking performed and a process time $X_{j2}(c)$ with reworking. This leads to:

$$X_j(c) = p_j(c) X_{j1}(c) + (1 - p_j(c)) X_{j2}(c) \quad (7)$$

where $p_j(c)$ is the probability for an individual body of color c to pass production step j . We assume the random variables to be independent. This is valid only because we neglect effects due to systematic errors like machine misconfigurations or color batches out of specification. From the layout of the paint shop and the distribution analysis of the processes we derive, that the processes X_1 and X_2 have no rework thus we set $p_1 = p_2 = 1$. Additionally we combine the two processes to one random variable Y_1 . In a second simplification we combine the three processes with

rework X_3, X_4, X_5 to one random variable Y_2 . Thus we get:

$$M(c) = m(c) + \sum_{k=1}^2 Y_k(c) \quad (8)$$

Obviously, the simplification reduces the number of unknown parameters of the color depending distributions significantly.

NUMERICAL RESULTS

In order to validate the various models, the real throughput times are compared with the throughput times of the models for every color c by calculating error sums.

To obtain the deviations the time is discretized and then the quadratic differences between the density of the data and the considered distribution $M(c)$ are summed. The most important error sums of the top-down analysis are shown in Table 1 (Danner, 2013).

	$M(c) \sim N(\mu, \sigma)$	$M(c) \sim GAM(\alpha, \lambda)$	$M(c) \sim LN(\mu, \sigma)$
mean	1,15	0,41	0,29
standard deviation	0,19	0,13	0,10
minimum	0,87	0,23	0,15
maximum	1,61	0,65	0,45

Table 1: Error sums of top-down analysis

As expected from the skewedness of the observed distributions, the normal distribution is not suitable for fitting throughput times of a paint shop. The gamma and the lognormal distribution on the other hand show promising results.

Figure 4 shows a typical histogram of the throughput time for one selected color. Also the fit with a lognormal distribution (Danner, 2013) is shown in this graph. The differences between the histogram and the lognormal distribution are very small.

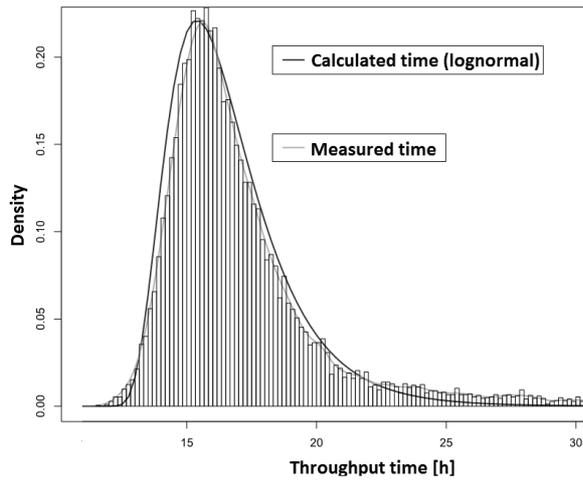


Figure 4: Throughput time of a paint shop fitted with a lognormal distribution

Next the procedure and the results of the bottom-up method are verified and validated. We assume the processes 1 and 2 to follow a normal or gamma distribution. We use the folding invariance of these distributions (Hübner, 2009) to yield a gamma function again. Note that we set the parameter λ to be equal for both gamma functions which is a prerequisite for this relation.

In Figure 5, the distribution of the throughput time for the process Y_1 is compared to the measurement. From this we conclude that there is no rework in processes 1 and 2. Thus the combination of the two processes is a valid simplification of the model in combination with the assumption of a normal or gamma distribution.

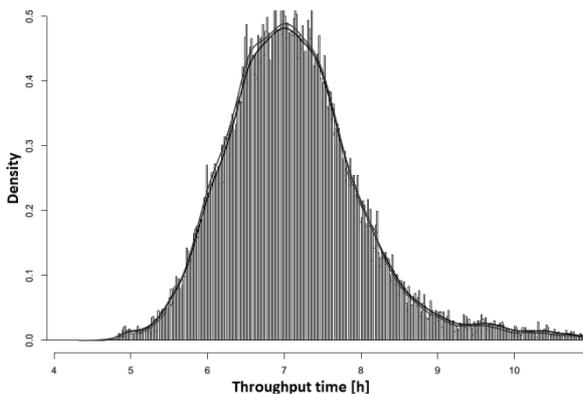


Figure 5: Histogram and density of throughput time Y_1

For the steps 3 to 5 (Y_2) the concept of folding of integrals to unify the three distributions into one cannot be applied (Danner, 2013). For logarithmic normal distributions the folding can result in other functions than the logarithmic normal distribution. For the Gamma function we cannot assume the parameter

λ to be equal for all three production steps due to the very different nature of the reworks. Nevertheless, we obtained good results by simplifying the last three processes to Y_2 and fitting with suitable distributions (Danner, 2013).

In Table 2, typical results or the error sums using the bottom-up method are shown (Danner, 2013).

	$Y_1 \sim N(\mu_1, \sigma_1),$ $Y_2 \sim LN(\mu_2, \sigma_2)$	$Y_1 \sim N(\mu, \sigma),$ $Y_2 \sim GAM(\alpha, \lambda)$	$Y_1 \sim GAM(\alpha, \lambda),$ $Y_2 \sim LN(\mu, \sigma)$
mean	0,12	0,23	0,17
standard deviation	0,08	0,09	0,07
minimum	0,07	0,14	0,09
maximum	0,36	0,50	0,36

Table 2: Error sums of bottom-up analysis

From studying various simulation models that have been created with the bottom-up analysis, we found that the most appropriate combination of distributions for the two sub processes Y_1, Y_2 is: $Y_1 \sim N(\mu_1, \sigma_1)$ and $Y_2 \sim LN(\mu_2, \sigma_2)$.

CONCLUSION

The correct description of the behavior of a paint shop forms the basis for an optimal simulation of modern automotive production. Precise reproduction of the processes is not recommended due to the complex structure and the volatility of the processes. Therefore, a methodology was developed to „measure“ the behavior of the paint shop using production data. In a top-down analysis of the throughput time a distribution was calculated based on modelling of these datasets using distribution functions. Here the log-normal distribution gives a good description. In a bottom-up analysis, the description was further refined. As a result from our data analysis we showed that a convolution of a normal and a log-normal distribution provides a good stochastic description for a paint shop.

REFERENCES

Arnold, D., Furmans, K., 2009. “Materialfluss in Logistiksystemen”. 6., expanded edition, Springer, Berlin, Heidelberg

Bayer, J., Collisi, T., Wenzel, S., 2003. "Simulation in der Automobilproduktion". Springer, Berlin, Heidelberg.

Curry, G. L., Feldman, R. M., 2011. "Manufacturing Systems Modeling and Analysis". Second Edition, Springer, Berlin, Heidelberg.

Danner C., 2013. "Optimales Scheduling unter Berücksichtigung stochastischer Wiederbeschaffungszeiten". master thesis at the University of Applied Sciences OTH Regensburg, Professorship of Informatics/Mathematics.

Hübner, G., 2009. "Stochastik". 5., improved edition, Springer, Berlin, Heidelberg.

Manitz, M., 2005. "Leistungsanalyse von Montagesystemen mit stochastischen Bearbeitungszeiten". Kölner Wissenschaftsverlag, Köln.

Meißner, S., 2009. „Logistische Stabilität in der automobilen Variantenfließfertigung“. Thesis at the TU München, Professorship of Materials Handling, Material Flow and Logistics.

Meyr, H., 2004. „Supply chain planning in the German automotive industry“, OR Spectrum (2004) 26: 447–470, Springer

Spieckermann, S., 2002. "Neue Lösungsansätze für ausgewählte Planungsprobleme in Automobilrohbau und -lackiererei". Shaker, Aachen.

Weyer, M., 2002. "Das Produktionssteuerungskonzept Perlenkette und dessen Kennzahlensystem". Helmesverlag, Karlsruhe.

AUTHOR BIOGRAPHIES

THOMAS HUSSLEIN was born in Munich, Germany and went to University of Regensburg, where he studied computational physics and obtained his degree in 1993. He obtained his PhD in physics from the University of Regensburg in 1996. Afterwards he did his Postdoc at the IBM Research Center Yorktown Heights and the University of Pennsylvania. In 1999 he founded OptWare in order to transfer mathematical methods into application which he is leading ever since. Since 2012 he has a teaching assignment for Operations Research at the University of Applied Sciences OTH Regensburg.

JÖRG BREIDBACH was born in Bendorf/Rhein, Germany and went to the University of Kaiserslautern

where he obtained his degree in physics in 1998. He got his PhD in physical chemistry in the field of numerical simulation of charge migration in molecules from the University of Heidelberg in 2006. In 2004 he joined OptWare where he is now heading the department for algorithmic development and optimization. Since 2009 he has a teaching assignment at the University of Applied Sciences OTH Regensburg.

MARKUS SEIDL was born in Regensburg, Germany. He studied Mathematics at the University of Applied Sciences OTH Regensburg and obtained his Diploma degree in 2006. Also in 2006 he joined OptWare where he is currently leading the group for product development.

CHRISTIAN DANNER was born in Regensburg, Germany. He studied Mathematics at the University of Applied Sciences OTH Regensburg and obtained his Master degree in 2013. In 2013 he joined OptWare as a member of the group for product development.

WOLFGANG LAUF was born in Kiel, Germany. He studied Mathematics at the Universities of Tübingen and Würzburg, where he obtained his PhD in mathematics in 1994. From 1998 to 2002 he gained a lot of experience in air traffic related optimization and simulation problems at Lufthansa. In 2002 he got a position as a professor for mathematics at the University of Applied Sciences OTH Regensburg. His main research interests are Operations Research, Function Theory and Financial Mathematics.

MULTI-CRITERIA APPROACH FOR EMERGENCY SERVICE ORDERS IN ELECTRIC UTILITIES

Vinicius Jacques Garcia, Daniel Pinheiro Bernardon
and Alzenira Abaide
Federal University of Santa Maria UFSM, Brazil.
Email: {viniciusjg, dpbernardon, alzenira}@ufsm.br

Julio Fonini
AES Sul - Power Utility, Brazil.
Email: Julio.Fonini@aes.com

KEYWORDS

Work orders, combinatorial optimization, power systems automation, vehicle routing, integer linear programming.

ABSTRACT

This paper proposes a multi-criteria approach to handle emergency orders under real-time conditions in electric power distribution utilities. It is described how the problem related to serve work orders in electric utilities is considered, with an aggregated objective function developed to handle the minimization of the waiting time for emergency services, the total distance travelled and the sum of all delays related to already assigned orders. After that, actual cases have shown the effectiveness of the proposed model to be adopted in real world applications.

INTRODUCTION

Electric power distribution utilities are charged of managing customer attendance and maintenance procedures in their network. The consideration of emergency scenarios makes the problem harder especially by assuming resource constraints (human and material) and strict regulation policies that establish targets and indices related to this context.

Considering that maintenance crews help to maintain the network under normal conditions, i.e., all the customers with power supply and non-technical problems associated with the electric network, emergency orders are normally related to equipment failures, overload conditions and interrupted conductors.

From this context, the most relevant aspect to be considered refers to the waiting time associated with the emergency orders, since the level of injury or danger of death imposes immediate response from the network operations center (NOC). The decision-making problem involves a considerable amount of data and several aspects and criteria, all of them related to network and equipment operation procedures. This context is close to that one described by (Ribeiro et al. 1995): “decision making is a process of selecting ‘sufficiently good’ alternatives or course of actions in a set of available possibilities, to attain one or several goals”.

Such a decision making process when referring to emergency services in electric utility generally involves not only the waiting time for emergency orders but also two even important aspects: the total distance traveled and the sum of all delays related to already assigned orders. The former sounds really intuitive, because the minimization of the total distance traveled by all crews improves their productivity by aggregating more time in their workday to complete the assigned orders. The latter aspect is that one associated with one contribution of this paper: the consideration of multitasked maintenance crews. They are always charged of pre-established routes that include orders known a priori when a set of emergency orders come up. This criterion of minimizing the sum of all delays represents the desired trade-off between the planning and actual scenarios, in such a way that they could be as similar as possible.

This paper proposes a multi-criteria mathematical model to handle emergency orders under real-time conditions. It comprises three criteria related to this problem: the minimization of the waiting time for emergency services, the total distance travelled and the sum of all delays related to already assigned orders.

This paper is organized as follows: first the emergency work order dispatch problem is described, followed by the corresponding mathematical model. After that, the heuristic approach, preliminary results and final remarks are presented.

PROBLEM DEFINITION

The emergency work order dispatch problem (EWODP) is carried out within 24 hours a day, 7 days a week, corresponding to a main task of the electric NOC. Assuming this non-stop period and the critical issues involved, a real time system may be suitable to assign a repair crew to each remaining emergency work order (EWO).

When developing a system able to assign one order to a given repair crew, the following goals must be assumed:

- Reducing the dispatch time;
- Improving network security on operation and maintenance procedures;
- Standardization of dispatch criteria in such a way they could be closely related to business process.

The main issue involved is the aim of reducing the average service time, which is defined as the sum of the waiting time, the travel time and of the order execution time. In this work we consider the decision problem of assigning an EWO to a given maintenance crew available, mainly focusing on the waiting time. The challenge comes from the business process usually adopted by utilities: they have multitasked repair crew generally in charge of commercial services (customer demand orders) when an EWO comes up. From this assumption follows specific characteristics that make the whole optimization problem some orders of magnitude greater in the sense of the complexity involved.

In this work it is described a problem that emerge from specific characteristics of route construction to meet customer demand in the context of an electric power distribution utility in Brazil, specifically with concern to the occurrence of EWO. The main inspiration for the analysis carried out to represent and solve the EWODP track its origin from the well-known traveling salesman problem (Lawler et. al. 1985) and its famous generalization: the vehicle routing problem (Toth and Vigo 2001).

In the considered utility, maintenance crews must execute a set of service orders, what remounts the construction of multiple routes. These crews have their start point in a depot that can be distinguished from each other and they do not need to return to their start point when the last service is completed.

The fundamental aspect that must be considered refers to the definition of several kinds of service orders, with high importance to the ones that are not known a priori. Two different sets can be defined: those orders known a priori and related to commercial services requested by customers and those orders that have their inherently aspect of emergency, which may occur at any moment. Every maintenance crew is able to execute these two kinds of orders.

When a maintenance crew begins its journey, its corresponding route to execute only those commercial orders known a priori is available. The occurrence of emergency scenarios imposes the most appropriate treatment in order to consider these EWOs that are coming up and have precedence over the commercial orders. Following the number and their corresponding geographical location of EWOs, one or more maintenance crews will be considered to complete these services and, as consequence, they will have their routes modified.

The problem that arises from this context is related to the need of restructuring the existing routes only populated by commercial orders, now including the pending EWOs in the beginning of each existing route. From this perspective, two scenarios may be assumed: (1) reprogramming the set of remaining commercial services of all maintenance crews; and (2) only inserting

the pending EWOs in the beginning of each route while maintaining unchanged the route related to commercial services. The first option has strict technological constraints since each maintenance crew receives a batch of orders to be executed when its journey starts and the communication to reprogram the route during the day may be a bottleneck by the existing status quo of telecommunication services in Brazil.

One important definition is related to the main goal of the problem. There are several objectives that can be assumed, including those conflicting ones. One of these is reducing the waiting time to execute emergency services, exactly by the risks associated with security of the electric power network.

Another objective, this one related to economical aspect, is reducing the total cost of routes, corresponding to the total time to complete all designed routes. In this case, both commercial and emergency are considered when calculating the cost. Even in this case it is already possible to note a conflict between cost and precedence of emergency services: the higher is the importance of emergency services, the greater will be the cost.

The third and last aspect to be considered refers to minimizing the sum of all delays related to the previously assigned services, in order to maintain the desired trade-off between the planning and actual scenarios.

Following these concepts and definitions, a mathematical model was developed as depicted below.

The mathematical model developed

The first assumption is that there is a given set of crews with their corresponding a priori routes, which include services called commercial orders. Given an instant of time in that a certain number of emergency orders come up, it is assumed that they will be assigned to the given crews in such a way that previous routes will not be changed. This fact will cause insertion of emergency services in the previous known routes, involving a decision of which subset will be assigned to each crew and in which position on the route.

On the following mathematical model, three criteria are used to integrate the objective function: the first, weighted by W_1 , corresponds to the latency cost of all emergency orders; the second, weighted by W_2 , includes the cost of all routes; the third, weighted by W_3 , aims to reduce the delay when considering the time when a commercial order i is completed and the end time of each route.

The following parameters are considered:

- 0 : dummy order to define the final destination point of every crew;
- V_e : set of emergency orders;
- V_c : set of commercial orders;

V_s	: set of start points, which represent the initial position of each crew;	$\sum_{\langle i,j,r \rangle \in E} x_{ijr} \leq 1$	$\forall i \in V_c \cup V_s, \forall r \in R$	(4)
V	: $V = V_e \cup V_c \cup V_s \cup \{0\}$	$\sum_{\langle i,j,r \rangle \in E} x_{ijr} \leq 1$	$\forall j \in V_c, \forall r \in R$	(5)
R	: set of routes / crews;			
t_0	: initial time for every crew;	$\sum_{\langle i,j,r \rangle \in E} x_{ijr} - \sum_{\langle j,l,r \rangle \in E} x_{jlr} = 0$	$\forall j \in V_e, \forall r \in R$	(6)
T	: end time for every crew's workday;	$\sum_{\langle i,j,r \rangle \in E} x_{ijr} - \sum_{h \in V_e} x_{h,suc(i),r} = 0$	$\forall i \in V_c \cup V_s, \forall r \in R$	(7)
$suc(i)$: the successor point of i in the a priori route, $i \in V_c$;	$t_i = t_0$	$\forall i \in V_s$	(8)
$pre(i)$: the antecessor point of i in the a priori route, $i \in V_c$;	$t_j \geq t_i + (c_{ij} + ts_j) +$ $+ \sum_{\langle i,j,r \rangle \in E} (x_{ijr} - 1)M$	$\forall j \in V, \forall i \in V$	(9)
$rC(i)$: the route index in which point i is inserted $i \in V_c$;	$t_i \geq t_{pre(i)} + c_{pre(i),i} + ts_i$	$\forall i \in V_c$	(10)
te_i	: time when the emergency request i came up, $i \in V_e$;	$t_i + ta_i - td_i = T$	$\forall i \in V_c$	(11)
ts_i	: execution time of order i , $i \in V \setminus \{0\}$;	$t_i - ts_i \leq T$	$\forall i \in V_e$	(12)
C	: cost related to each non-assigned emergency order;	$t_i - ts_i \geq te_i$	$\forall i \in V_e$	(13)
E	: $E = \{\langle i,j,r \rangle; i \in V_e, j \in V, r \in R,$ $r = rC(j)\} \cup \{\langle i,j,r \rangle; i \in V,$ $j \in V_e, r \in R, r = rC(i)\} \cup \{\langle i,j,r \rangle;$ $i \in V_e, j \in V_e, r \in R, i \neq j\} \cup$ $\{\langle i,0,r \rangle; i \in V_e, j \in V, r \in R\}$	$ta_i \geq 0$	$\forall i \in V_c$	(14)
$c_{i,j}$: travel time between points i and j ;	$td_i \geq 0$	$\forall i \in V_c$	(15)
M	: a huge value, typically $2T$;	$t_i \geq 0$	$\forall i \in V$	(16)
$W1, W2$: weighted factors of each objective function			
$, W3$	component, with $W1+W2+W3=1$.			

The following decision variables are defined:

$$x_{ijr} \begin{cases} 1 & \text{if point } j \text{ is successor of point } i \text{ in the route } r; \\ 0 & \text{otherwise;} \end{cases} \quad y_i \in \{0,1\} \quad \forall i \in V_e \quad (17)$$

$$y_i \begin{cases} 0 & \text{if the emergency order } i \text{ is assigned to some} \\ & \text{route;} \\ 1 & \text{otherwise;} \end{cases} \quad x_{ijr} \in \{0,1\} \quad \forall \langle i,j,r \rangle \in E \quad (18)$$

t_i : time when order i is completed;

ta_i : $ta_i = T - t_i$ if $t_i < T$; 0 otherwise;

td_i : $td_i = t_i - T$ if $t_i > T$; 0 otherwise;

$$\text{Min} \quad W_1 \sum_{i \in V_e} t_i + W_2 \sum_{\langle i,j,r \rangle \in E, i \in V_e} c_{ij} x_{ijr} + W_3 \sum_{i \in V_c} td_i \quad (1)$$

Subject to:

$$\sum_{\langle i,j,r \rangle \in E} x_{ijr} + y_i = 1 \quad \forall i \in V_e \quad (2)$$

$$\sum_{\langle i,j,r \rangle \in E} x_{ijr} + y_j = 1 \quad \forall j \in V_e \quad (3)$$

COMPUTATIONAL METHODOLOGY PROPOSED

When observing the literature, several contributions explain how to consider the vehicle routing problem in a context much similar to that one defined in this paper.

(Okonjo-Adigwe 1988) proposes a method to generate balanced routes to the vehicle routing problem that includes upper and lower bounds on the route time of each vehicle, obtained by a heuristic algorithm. After that and incorporating these limits, one mathematical model is derived and the problem is optimally solved. (Chandran et. al. 2006) have developed an approach to have balanced workload by modeling the vehicle routing problem first as a clustering problem, in a classical cluster-first, route-second approach. (Weintraub et. al. 1999) consider the emergency vehicle

dispatching problem and the workload balanced is obtained by a post-optimization procedure that includes order interchange between routes, clustering and routing.

(Anbuudayasankar et. al. 2009) have pointed out that the workload balanced should not be assumed in the sense of total route cost but in the sense of equity when considering dangerous and strenuous activities.

One can note that the goal of balancing routes has strict relation with human relations in the company, since it is the most apparent aspect that is evaluated when comparing the work effort between two given crews. In this work, the considerations of (Anbuudayasankar et. al. 2009) are included in the form of priority and execution time of each service, what allows minimizing and balancing the total route time.

The methodology to solve the EWODP proposed in this paper comprises a mixed linear programming mathematical model to be included in a computational system, which is able to execute a real-time automatic dispatch of EWOs. However, real-time conditions, strict constraints related to workday of every crew and even pre-assigned orders endow a certain degree of complexity to the problem of this approach.

The previous defined mathematical model can be used to solve “small” instances (i.e., with less than 20 orders) to optimality. Actual instances, mainly by assuming the proposal of applying this methodology to actual scenarios, may typically involve more than 100 orders to be assigned to 4 to 10 crews.

A heuristic procedure may be suitable and convenient to this context, and exactly this root was followed by the computational methodology proposed in this paper.

```

IterativeConstructDeconstruct( $V, E, R, T, c, W, N$ )
1.  $S = \text{Construct}(0, V, E, R, T, c, W)$ ;
2.  $i = 0$ ;
3.  $S = \text{evaluate}(S, V, E, T, c, W)$ ;
4. while( $i < N$ ) do
5.      $Sd = \text{Deconstruct}(S, V, E, R, T, c, W)$ ;
6.      $Sd = \text{evaluate}(Sd, V, E, T, c, W)$ ;
7.     if ( $Sd < S$ ) then  $S = Sd$ ;
8.      $i = i + 1$ ;
9. return( $S$ );

```

Figure 1: Algorithm proposed to solve the EWODP.

The procedure of Figure 1 is mainly inspired by (Ahmadi and Osman 2004) and includes two main phases: the former, charged of construct a solution; and the latter that deconstructs the previous assignment of “construct” phase. Since every crew route will be formed by a sequence of orders, the EWODP should define essentially a sequence of orders including the emergency ones. In “construct” phase, a Monte Carlo method is conducted by defining this sequence, according to the objective function of equation (1). In

“deconstruct” phase, all emergency orders are extracted from the sequence at random and included in the best position according to the objective function of equation (1). This procedure is repeated by N iterations, always keeping the best solution found. Parameters V, E, R, T, c and W are exactly the same described in the previous section as part of the mathematical model described.

PRELIMINARY RESULTS

In order to evaluate how suitable may be the algorithm developed, preliminary results were obtained when the system approaching the following actual case:

- 25 commercial orders;
- 3 repair crews;
- 5 emergency work orders (EWO).

Figure 2 shows 3 repair crews and their corresponding routes, namely 3056, 4019, and 4025. All of them begin their work at 10 am and finish their journey before 3:10pm. The scenario of Figure 2 shows that every repair crew is already charged of commercial orders when it will serve EWOs. There are two main charts on this figure: the former points out how much of the total time of each crew will be on attendance services and how much time will be spent on displacement; the latter is the timeline for each crew, where the black color refers to emergency orders, red color refers to orders with priority p0, yellow color refers to orders with priority p1 and green color refers to orders with priority p2. This interface is of Google Earth platform (Google Earth 2013) that shows the result of the developed algorithm after describing a kml file.

From scenario of Figure 2, Table 1 shows the schedule for each maintenance crew only considering commercial orders previously assigned to. There will also be three kinds of priority levels for all commercial orders: lower will be the most critical service and all routes have their first part formed by orders of level 0 (p0), followed by orders of level 1 (p1) and finally orders of level 2 (p2).

Table 1: Schedule for commercial orders.

#Crew	Start time	Finish time	#orders p0	#orders p1	#orders p2
3056	10:00am	3:06pm	3	5	4
4019	10:00am	2:33pm	3	3	2
4025	10:00am	12:22pm	0	4	1

At 10:10 am, 5 EWOs come up and three available crews must attend them. This aspect conducts a new routing for all crews in order to minimize the function of equation (1), whose result is shown in Figure 3.

It can be noted in Figure 3 that there is more time spent in displacement when comparing the Team’profile of Figure 3 with Team’s profile of Figure 2. That behavior is easy to understand since these 5 EWOs are included

in the beginning of all routes, as depicted in Table 2. One important aspect refers to the maintenance of the previous assigned routes: for all routes, it was maintained the commercial orders assigned, just causing

a delay on the finish time due to attendance of emergency orders.

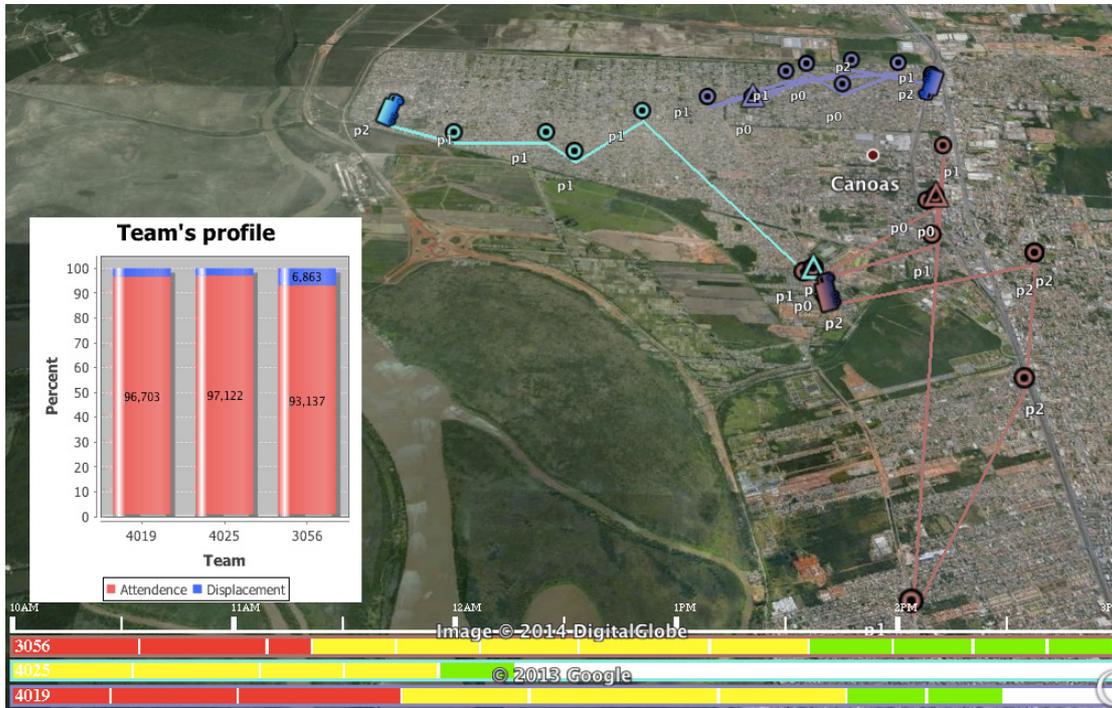


Figure 2: Maintenance crew routes for commercial orders.

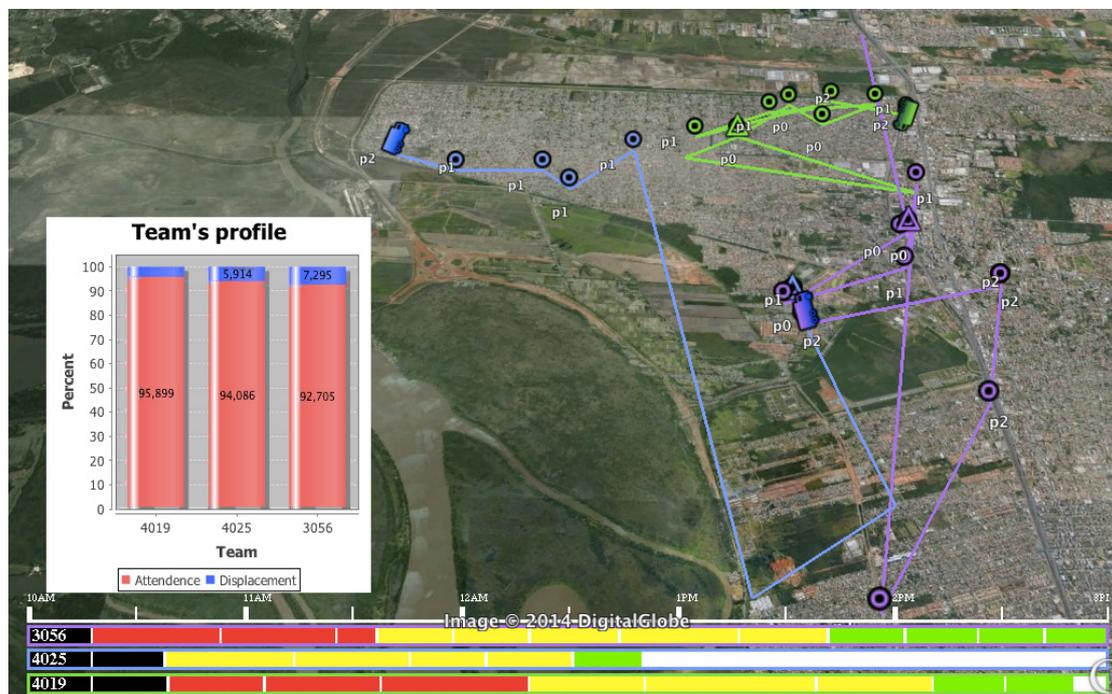


Figure 3: Maintenance crew routes after considering 5 EWOs.

Table 2: Schedule for both commercial and emergency orders.

#Crew	Start time	Finish time	#emerg. orders	#order s	#orders p1	#orders p2
3056	10:00am	3:32pm	1	3	5	4
4019	10:00am	3:19pm	2	3	3	2
4025	10:00am	1:08pm	2	0	4	1

FINAL REMARKS

This paper has presented a computational methodology to address the emergency work order dispatch problem. One key aspect involved in this system corresponds to the real-time condition, leading to development of an algorithm with response times in the order of microseconds or at least milliseconds.

Considering the main contributions of this paper, the following must be mentioned:

1. Reducing on the dispatch time;
2. Improving network security on operation and maintenance procedures;
3. Standardization of dispatch criteria in such a way that they could be closely related to the business processes adopted.

The benefits from adopting this approach may refer not only to the capability of managing in a secure manner high critical tasks but also by making possible a great amount of calculations and analysis required in the decision making process.

Future works should address the development of alternative mathematical models to this problem.

ACKNOWLEDGEMENTS

The authors would like to thank the AES SUL Distribuidora Gaúcha de Energia SA for financial support provided to the project "Sistema de apoio à decisão para despacho automático e integrado de ordens de serviço emergenciais".

REFERENCES

- Ahmadi, S., Osman, I. H. Density Based Problem Space Search for the Capacitated Clustering. *Annals of Operations Research*, 131, 21-43, 2004.
- Anbuudayasankar, S. P., Ganesh, K., Lenny Koh, S. C., Mohandas, K. Clustering-based heuristic for the workload balancing problem in enterprise logistics. *Int. J. Value Chain Management*, 3(3), 302-315, 2009.
- Chandran, N., Narendran, T. T., Ganesh, K. A clustering approach to solve the multiple travelling salesmen problem. *Int. J. Industrial and Systems Engineering*, 1(3), 372-387, 2006.

Google Inc. Google Earth (Version 7.0.3.8542) [Software]. Available from <http://www.google.com/earth/download/>, 2013.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., Shmoys, D. B. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*, Wiley, 1985.

Okonjo-Adigwe, C. An effective method of balancing the workload amongst salesmen. *Omega*, 16(2), 159-163, 1988.

Ribeiro, R. A., Powell, P. L. e Baldwin, J. F. Uncertainty in decision-making: An abductive perspective, *Decision Support Systems* 13: 183-193, 1995.

Toth, P., Vigo, D. *The Vehicle Routing Problem Discrete Math*, Siam Monographs on Discrete Mathematics and Applications, 2001.

Weintraub, A., Aboud J., Fernandez C., Laporte, G., Ramirez E. An emergency vehicle dispatching system for an electric utility in Chile. *Journal of the Operational Research Society*, 50, 690-696, 1999.

AUTHOR BIOGRAPHIES

VÍNÍCIUS JACQUES GARCIA was born in Santo Ângelo, Brazil, in April 1976. He received his Bachelor's degree from the Federal University of Santa Maria in 2000, the Master's and Doctor's degree from the State University of Campinas in 2002 and 2005, respectively. Since 2011 he has been professor at Federal University of Santa Maria. His research interests include distribution system planning and operation, combinatorial optimization and operations research.

DANIEL PINHEIRO BERNARDON is a professor of power systems at Federal University of Santa Maria. His research interests include smart grid, distributed generation, distribution system analysis, planning and operation.

ALZENIRA DA ROSA ABAIDE was born in Santa Maria, Brazil. He received the Bachelor's degree from the University of Santa Maria in 1980, the Master's (2000) and the Post-Graduated as Doctor in Electric Engineer (2005). She worked as engineer from 1986 to 1988 in State Company of Electric Energy. She has been professor and researcher at Federal University of Santa Maria since 1989.

JÚLIO FONINI received his Bachelor's degree from the University of Vale do Rio dos Sinos in 2013. Currently, he is a production engineer at AES Sul Power Utility, in Brazil.

AN IMPROVED RECEDING HORIZON GENETIC ALGORITHM FOR THE TUG FLEET OPTIMISATION PROBLEM

Robin T. Bye and Hans G. Schaathun
Faculty of Engineering and Natural Sciences
Aalesund University College
Postboks 1517, NO-6025 Ålesund, Norway
Email: {roby, hasc}@hials.no

KEYWORDS

Dynamic resource allocation; genetic algorithm; cost functions; receding horizon control; tug fleet optimisation simulator.

ABSTRACT

A fleet of tugs along the northern Norwegian coast must be dynamically positioned to minimise the risk of oil tanker drifting accidents. We have previously presented a receding horizon genetic algorithm (RHGA) for solving this tug fleet optimisation (TFO) problem. In this paper, we begin by presenting an overview of the TFO problem and the details of the RHGA. Next, we identify and correct a flaw in the original cost function of the RHGA. In addition, we present several new cost functions that can be used for dynamic resource allocation by an algorithm such as the RHGA. In a preliminary simulation study, we correct and extend the simulation scenarios used in our previous work and examine the merit of each of the suggested cost functions. Finally, we discuss the potential for an objective evaluation method for comparing various TFO algorithms and briefly present our TFO simulator.

INTRODUCTION

Thousands of ships, including several hundred oil tankers, move along the northern Norwegian coastline every year, making it susceptible to the risk of drift grounding accidents and oil spill (Havforskningsinstituttet, 2010). Constantly attempting to reduce the risk of such accidents, the Norwegian Coastal Administration (NCA) runs a vessel traffic services (VTS) centre in the town of Vardø that administers a fleet of tugs patrolling the coastline. The role of the VTS centre is to continuously order the tugs to new positions in a manner such that if an oil tanker loses manoeuvrability, e.g., through steering or propulsion failure, there will be at least one tug sufficiently close that it can intercept the drifting oil tanker before it runs ashore (Eide et al., 2007a).

To aid the NCA with positioning their fleet of tugs, a set of risk-based decision support tools based on dynamical risk models have been developed (Eide et al., 2007a,b). The risk models are based on real-time information such as wind, waves, currents, geography, kind of oil tankers in transit, their crew, and the estimated oil spill size and potential impact, to mention some. Using the decision

support tools aids the human operator at a VTS centre in directing tugs towards high-risk target areas.

The abovementioned decision support tools do not tell explicitly which tugs should move where; that is still an informed decision based on the operators experience and currently available information. Since the number of oil tanker transits are expected to rise significantly in coming years (Havforskningsinstituttet, 2010), the problem of positioning the tugs can quickly grow and become unmanageable for human operators. Consequently, there is a need of an algorithm able to calculate position trajectories that each tug should follow in order to reduce the overall risk of drifting accidents.

In our Dynamic Resource Allocation with Maritime Application (DRAMA) research group at the Aalesund University College (AAUC), we have solved this tug fleet optimisation (TFO) problem by means of a receding horizon genetic algorithm (RHGA) (Bye et al., 2010; Bye, 2012). This algorithm combines methods from control theory and computational intelligence to iteratively plan movement trajectories for each individual tug such that the net collective behaviour of the tugs as measured by a cost function is optimised by means of a genetic algorithm (GA). Subsequently, at last year's meeting of this conference, we presented a modified version of the RHGA called the receding horizon mixed integer programming algorithm (RHMIPA) in which we reformulated our choice of cost function such that it turned into a linear programming problem (Assimizele et al., 2013). Notably, the cost function is the same in the RHGA and the RHMIPA, it is only the mathematical formulation that differs. Whereas the RHGA typically will return a good, albeit inexact and suboptimal solution, at every run, the RHMIPA in contrast finds an exact, global minimum of the cost function.

Since both algorithms are identical except for the method used for minimising the very same cost function, a simple measure for comparison is simply the accumulated cost, for which the exact MIP solver in the RHMIPA will cause it to outperform the RHGA, which uses a suboptimal, heuristic GA solver. Unsurprisingly, perhaps, this superiority of the RHMIPA comes at the cost of slower computational evaluation when compared to the heuristic RHGA (Assimizele et al., 2013).

In keeping with the real-world nature of the TFO problem, both algorithms were also compared with a realistic

option that the NCA and Norwegian policymakers regularly consider, namely that of a static policy in which the tugs are uniformly positioned at base stations spread out along the coastline. Obviously, in terms of cost function minimisation, the active tug fleet patrol scheme of the RHGA and the RHMIPA both outperform the static policy in which tugs are waiting passively for an incident to occur (Bye et al., 2010; Bye, 2012; Assimizele et al., 2013).

The work presented in this paper was motivated by the desire to (1) continue our development of the RHGA with a particular focus on optional cost functions, and (2) rewrite and formalise the implementation of the algorithm and simulation scenarios in a TFO simulator.

With respect to (1), we recently identified what appears to be a flaw in the cost function used for the RHGA and RHMIPA. In addition to rectifying this error, we wanted to examine several other choices of cost functions. A particular challenge, then, is the comparison and evaluation of cost functions. Using a naive heuristic or a simple method such as the static policy above constitute an indirect measure for comparing algorithms, where each algorithm's performance versus the naive heuristic or static policy is compared instead of a direct comparison of the algorithms' ability to minimise some common cost function. However, when designing new algorithms for the TFO problem that employ new and different cost functions, the methods for comparing algorithms above are of limited value, since different cost functions by definition are not directly comparable. Hence, there is a need for some kind of common, objective method for evaluating the merit of a particular choice of cost function in a TFO algorithm.

With respect to (2), our previous work has shown us that the TFO problem is an excellent case study for the DRAMA research group, with still many aspects of the TFO problem yet to be examined. Specifically, we want to investigate how a TFO algorithm can be able to handle a variety of simulation scenarios, including oil tankers entering or leaving the patrol zone; changes in number of tugs and oil tankers; changing weather conditions, drift trajectories, and maximum tug speeds; and much more (see Discussion). To answer these questions, we have completely rewritten our code base using the advanced, purely-functional programming language Haskell.¹ Part of the motivation for choosing a functional language like Haskell was to enable fast prototyping while keeping our code robust, concise, and not the least correct. Another reason was the potential for extensions into parallel programming, which may be required as the simulator grows more complex and more computational resources are needed.

In the following sections, we proceed by presenting the formulation of the TFO problem as defined in Bye et al. (2010); Bye (2012); Assimizele et al. (2013). Next, we point out what we believe is a flaw in the original cost function employed in the RHGA, and present a number of optional cost functions. We then propose a new and objective evaluation method that can be used for

comparing various TFO algorithms. Finally, we present some simulation results and discuss the viability of our approach as well as future work.

TFO PROBLEM FORMULATION

For formulating the TFO problem, we adopt most of the assumption in our previous work (Bye et al., 2010; Bye, 2012; Assimizele et al., 2013). First, assume that N_o oil tankers move in one dimension only (north or south, say) along a line of motion z . This is a reasonable assumption considering that oil tankers by law follow predefined piecewise-linear corridors. Second, inside of z and closer to shore, assume that N_p tugs are patrolling along a line of motion y parallel to z . Although the coastline is rather rugged, with fjords, peninsulas, and islands, tugs should stop drifting ships before they reach land or danger zones, thus a straight patrol line some distance from the rugged coastline can be considered a conservative choice.

Next, we assume real-time access to simulation data from a set of accurate models able to predict future positions of oil tankers along z and the corresponding potential drift trajectories given current and predicted information about the tankers themselves and the environment they are operating in. Such models exist and are currently an active focus of research (e.g., see Hackett et al. (2006); Breivik and Allen (2008); Breivik et al. (2011)).

For example, consider an oil tanker currently positioned at $z(t)$. There is a small chance that the tanker may suffer from engine failure or some other incident and start drifting right now at $t = t_d$. However, if not, it may also continue sailing along z . We may predict the future positions of the tanker some time T_h hours ahead in time, where T_h is called the prediction horizon. Employing a discrete-time model with a sampling period of $T_s = 1$ hour, the estimated future tanker positions are given by $\{\hat{z}(t|t_d)\}$ for $t = t_d + 1, t_d + 2, \dots, t_d + T_h$.

For each predicted point $\hat{z}(t|t_d)$, there is a corresponding predicted drift trajectory starting at $\hat{z}(t|t_d)$ that may or may not intersect the patrol line y after an estimated drift time $\hat{\Delta}$ into the future depending on ocean currents, wave heights, wind conditions, oil tanker shape and weight, and other factors.

In previous work, we either set the estimated drift time $\hat{\Delta}$ to be 8 hours for all oil tankers (Assimizele et al., 2013), or to be drawn randomly for each oil tanker from a uniform probability distribution in the interval $[8, \dots, 12]$ hours (Bye et al., 2010; Bye, 2012). According to Eide et al. (2007a), these drift times correspond to situations of "fast drift" and not the typical, or average, case. On the other hand, it should be kept in mind that there will inevitably be a delay between when an oil tanker begins drifting and when the VTS centre actually is notified of the incident and can order tugs to the rescue.

Collecting all predicted drift trajectories for all oil tankers results in a distribution of *crosspoints* located at points where future drift trajectories will intersect the patrol line y . A crosspoint of the c th oil tanker's drift trajectory at time t can be defined as y_t^c . Taking the drift

¹<http://www.haskell.org>

time $\hat{\Delta}$ into account, a drift trajectory starting on z at $t = t_d$ will have a cross point on y at $t = t_d + \hat{\Delta}$. Assuming the same drift time for all drift trajectories and considering the prediction horizon T_h , there is a predicted set of crosspoints given by

$$\{y_t^c\} = \{y_{t_d+\hat{\Delta}}^c, y_{t_d+1+\hat{\Delta}}^c, \dots, y_{t_d+T_h}^c\} \quad (1)$$

In addition to crosspoints, we define a *patrol point* (tug position on y) on the p th tug's patrol trajectory at time t as y_t^p .

Based on the predicted future distribution of crosspoints, the TFO problem is to calculate trajectories, or sequences of patrol points, along y for each of the patrolling tugs such that the risk of an oil tanker in drift not being reached and prevented from grounding is minimised.

Figure 1 shows a graphical representation of the problem description, exemplified by two patrolling tugs and three oil tankers.

COST FUNCTIONS

Original Cost Function f_1

Determining a suitable cost function for optimisation algorithms such as the RHGA and the RHMIPA is imperative for the algorithm to be able to find desirable solutions. The cost function we present firstly is the same as the one used in these algorithms (Bye et al., 2010; Bye, 2012; Assimizele et al., 2013) and is defined as the sum of the distances between all crosspoints and the *nearest* patrol points. The rationale behind this choice is that if an oil tanker in drift can/cannot be saved by a tug some distance away, it is not important that other tugs further away can/cannot save it at a later time.

For N_o oil tankers and N_p patrol tugs, the cost $f_1(t)$ is defined mathematically as

$$f_1(t) = \sum_{t=t_d}^{t_d+T_h} \sum_{o \in O} \min_{p \in P} |y_t^c - y_t^p| \quad (2)$$

for each oil tanker $o \in O = \{1, 2, \dots, N_o\}$ and each patrol tug $p \in P = \{1, 2, \dots, N_p\}$.

Note that choosing distance as a cost measure is equivalent to minimum rescue time if one assumes that all tugs have the same maximum speed. For cases where tugs have different maximum speeds, one could define rescue time as distance divided by maximum tug speed and add up the minimum rescue times for each cross point.

An example scenario with six oil tankers and three tugs is shown in Figure 2 (adapted from (Assimizele et al., 2013)), where an optimal solution found by the RHMIPA (bottom) is compared with a static policy (top) where tugs simply remain at their individual base station. Employing the RHMIPA, the patrol tugs spread out and track different clusters of crosspoints, thus collectively reducing the overall risk of grounding.

Cost Function f_2

The cost function f_1 presented above adds up the absolute value of the distance between every cross point and its

nearest patrol point. This means that in situations where a particular patrol point lies between two crosspoints, its exact position does not affect the cost, since being closer to one of the crosspoints means being further away from the other. This may or may not be what we want. If we prefer the patrol point to be positioned midway between the two crosspoints, we could use the square of the distance instead of the absolute value, such as in cost function f_2 :

$$f_2(t) = \sum_{t=t_d}^{t_d+T_h} \sum_{o \in O} \min_{p \in P} |y_t^c - y_t^p|^2 \quad (3)$$

The reason is that by using the square, we punish larger distances more than smaller distances.

A Flaw In The Original Cost Function

Let the alarm time t_a denote when the tugs are alarmed that an oil tanker is adrift, and, as before, let t_d be the time the oil tanker actually starts drifting. Using the cost functions f_1 and f_2 for planning the trajectories of the tugs implies the assumption that $t_a = t_d$, that is, the tugs are alarmed immediately, and that tugs will continue to execute their original plans even after receiving an alarm. In reality, however, these assumptions are unrealistic. First of all, oil tankers will typically have drifted for some time, 3 hours say, before the tugs are alarmed, and hence, in general, t_d will occur earlier than t_a . Consequently, we can define a new, and shorter, drift-from-alarm (DFA) time $\hat{\Delta}_a$, which is the drift time from the tugs receive an alarm at t_a until the drifting tanker crosses the patrol line at a crosspoint. For example, let us consider Figure 2, and assume that all oil tankers will take an estimated $\hat{\Delta} = 11$ hours, say, to drift aground from current positions. Hence, the first crosspoints that appear at $t = 8$ correspond to oil tankers starting drifting at $t = t_d = -3$, the NCA being alarmed 3 hours later at $t = t_a = 0$, and the DFA time becomes $\hat{\Delta}_a = 8$ hours. Likewise, the crosspoints that appear at $t = 9$ correspond to oil tankers starting drifting at $t = t_d = -2$, the NCA being alarmed at $t = t_a = 1$, the DFA time becomes $\hat{\Delta}_a = 8$ hours, and so on.

Second, when alarmed, tugs should abandon their plans and make every effort to intercept a drifting tanker before it runs aground. More relevant, therefore, are the positions of the tugs when they receive the alarm at time t_a , and the hypothetical future positions where drifting tankers will cross the patrol line some $\hat{\Delta}_a$ hours later, where $\hat{\Delta}_a$ is the total drift time (8–12 hours) less the time it takes before the tugs are being alarmed (3 hours), thus $\hat{\Delta}_a$ is in the range 5–9 hours.

Cost Function f_3

To address the issues raised above regarding the original cost function f_1 , we propose a modified cost function f_3 as given below:

$$f_3(t) = \sum_{t=t_a}^{t_a+T_h} \sum_{o \in O} \min_{p \in P} |y_{t+\hat{\Delta}_a}^c - y_t^p| \quad (4)$$

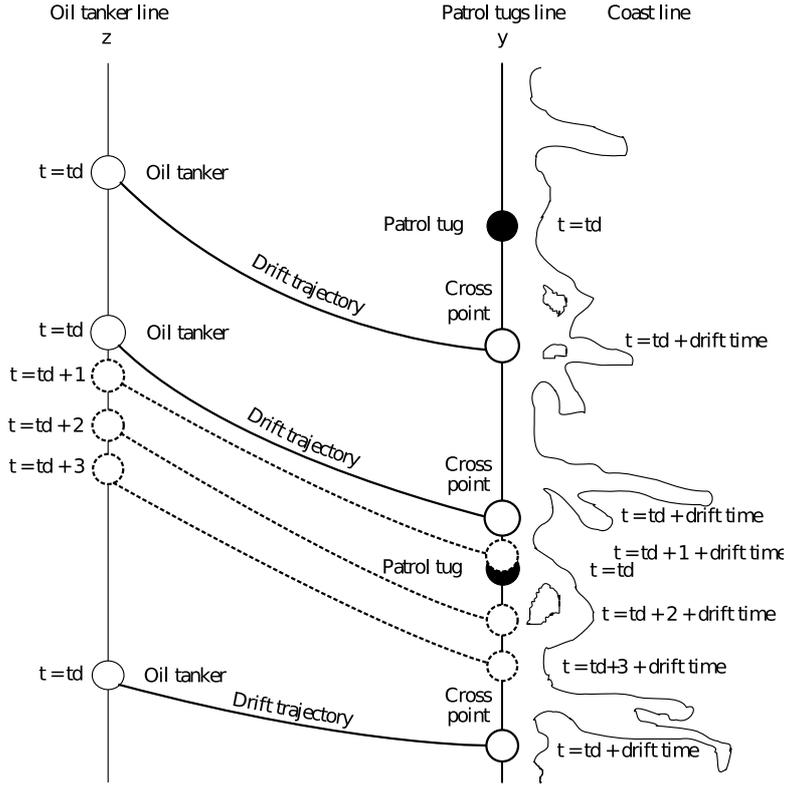


Figure 1: TFO Problem: Where Should Tugs Move?

Compared with f_1 , f_3 is modified in two ways: (1) in the first sum, time t is upper-limited by $t_a + T_h$ instead of $t_d + T_h$ and lower-limited by t_a instead of t_d ; and (2) in the absolute value term, we measure the distance between each cross point at some future cross time $t + \hat{\Delta}_a$ and the position of the nearest tug at the alarm time t , and do this for the current and future potential alarm times.

Cost Function f_4

Again, combining the option of squaring the distances in f_2 with the modification in f_3 , we also propose the cost function f_4 given by

$$f_4(t) = \sum_{t=t_a}^{t_a+T_h} \sum_{o \in O} \min_{p \in P} |y_{t+\hat{\Delta}_a}^c - y_t^p|^2 \quad (5)$$

Cost Function f_5

Yet another option for the choice of cost function is to categorise crosspoints within a certain safe range r as very likely to be reachable before grounding by a particular tug and therefore not to include these crosspoints in the cost function evaluation. Considering f_3 presented above, we simply subtract the safe range r from the distance and if the result is negative, we raise it to zero, as shown in f_5 below:

$$f_5(t) = \sum_{t=t_a}^{t_a+T_h} \sum_{o \in O} \max \left\{ 0, \min_{p \in P} |y_{t+\hat{\Delta}_a}^c - y_t^p| - r \right\} \quad (6)$$

A reasonable and conservative choice for r could for instance be half the expected distance a tug can travel from

an alarm is received until the first hypothetical crosspoints occur.

In terms of minimising this cost function, one challenge will be that of flat cost surface regions for crosspoints within the safe range, which makes it more difficult to find an optimal solution.

Cost Function f_6

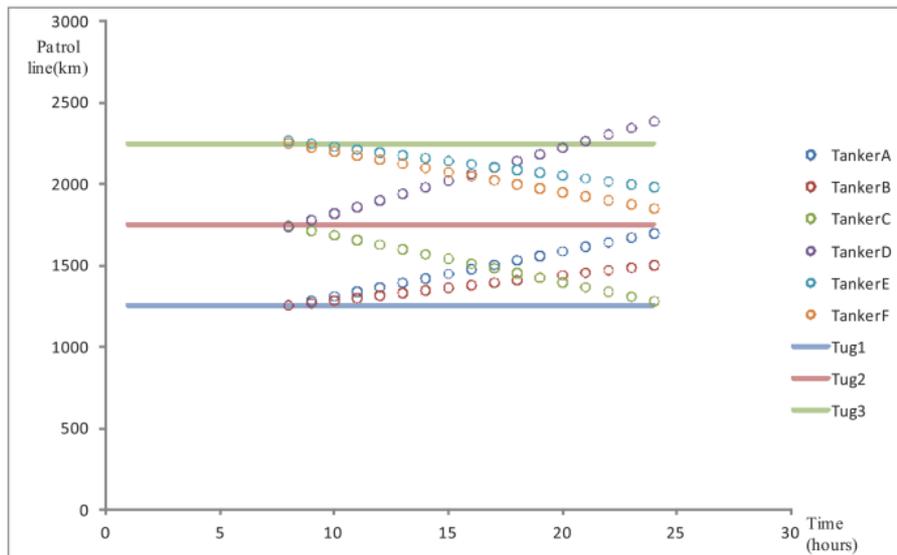
For the last cost function we will present, we continue to use the safe range r for measuring the number of unreachable crosspoints. If crosspoints are outside the safe range, we add 1 to the accumulated cost, otherwise 0. The cost function is given by f_6 below:

$$f_6(t) = \sum_{t=t_a}^{t_a+T_h} \sum_{o \in O} g \left(\min_{p \in P} |y_{t+\hat{\Delta}_a}^c - y_t^p| \right), \quad (7)$$

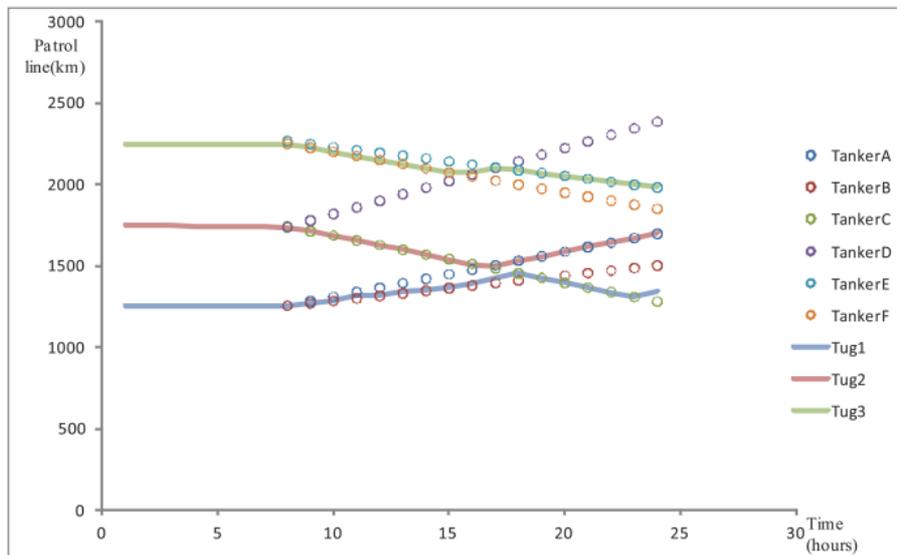
$$g(x) = \begin{cases} 1, & \text{if } x > r. \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Objective Evaluation Method For TFO Algorithms

By definition, the costs of different cost functions are not directly comparable, and we therefore need some kind of objective evaluation method for making comparisons. Here, we suggest that one such method is to (1) generate a deterministic and reproducible simulation scenario; (2) run the RHGA (or another TFO algorithm) for a given number of planning steps; (3) considering each oil tanker separately, assume each tanker begins drifting and count the number of salvageable tankers; (4) for the same simulation scenario, repeat (2) and (3) with a different



(a) Static Policy



(b) Optimal RHMIPA Solution

Figure 2: Example TFO Scenario

cost function in the RHGA (or a different TFO algorithm); and (5) repeat (1)–(4) for a number of different simulation scenarios and find the accumulated objective evaluation cost for each cost function (or TFO algorithm).

A simulation scenario in this case is simply a set of pre-determined oil tanker movements and the resulting hypothetical drift trajectories and crosspoints for a pre-specified duration. For testing purposes, we can generate a number of such scenarios offline and use them as input data for testing TFO algorithms. In a real-world application, the actual scenario is what that is happening right now, and future oil tanker positions, drift trajectories, and crosspoints would have to be predicted in real-time.

Other possible objective measures exist, e.g., we could sum up the total fuel consumption and use it as a component of an overall objective measure if that is of interest.

RECEDING HORIZON CONTROL

Above, we have presented a number of cost functions that can be minimised, with a GA or otherwise, to return a set of optimal tug trajectories. However, we need to present a method for handling the dynamic nature of the environment and parameters involved, and therefore adopt the principle of receding horizon control (RHC) that we have employed previously Bye et al. (2010); Bye (2012). Because neither oil tankers' speed and heading, nor wind, wave, and ocean current conditions are static, the resulting predicted future distribution of crosspoints will change, and patrol trajectories optimised by the GA will soon become outdated. One possibility for overcoming this problem is to run the GA at regular intervals, constantly incorporating updated *current* information about the state of the oil tankers and weather conditions as well as updated *predictions* of these factors. While tugs begin to move according to the solutions planned by the GA,

new patrol trajectories can be calculated and replace the old ones. This feedback strategy is equivalent to a RHC scheme, which is interchangeably termed model predictive control (MPC) in the literature (e.g., see Maciejowski (2002); Rossiter (2004) for theoretical treatments).

In RHC, a control strategy that minimises some cost function is calculated a prespecified duration, namely the prediction horizon, into the future. However, only the first portion of this strategy is implemented before another control strategy is calculated based on new and predicted information available. The new solution replaces the old one but again only the first portion is implemented. This process repeats as a sequence of RHC planning steps.

A particular advantage of using RHC is that constraints can be handled in the design phase and not post hoc (e.g., see Goodwin et al. (2001); Maciejowski (2002)). For tugs, one such constraint is the inherent limitation of moving no faster than the maximum possible speed limited by the ship's engine, weather conditions, or even the wish to save fuel if one wants to take that into account. This maximum speed limits the number of reachable crosspoints. Using RHC it is possible to incorporate this constraint in the planning of tug patrol trajectories.

GA Optimisation Between Planning Steps

In the RHGA, a GA is used to solve an optimisation problem at every RHC planning step. A good choice of initial population allows the GA to find good solutions in fewer iterations than simply using a random population. It is possible to take the dynamics of the simulated scenario into account and, assuming that the scenario will not change significantly, a solution found at one planning step should also be a viable solution at the next planning step. This is achieved by an elitist strategy of keeping (a slightly modified version of) the best chromosome at one RHC step and inserting it into the initial population of the GA at the next RHC step. More details on the RHGA is outside the scope of this paper and has been presented previously (Bye et al., 2010; Bye, 2012).

SIMULATION RESULTS

Figure 3 shows an example simulation scenario where three tugs (black circles) are positioned at $y = [-500, 0, 500]$ at $t = 0$ and six oil tankers (not shown) are randomly positioned in open water along the oil tanker corridor z , limited to an observation zone of $[-750, 750]$ km. All plots depict time along the horizontal axis and position along the vertical axis. Using each of the cost functions presented previously, the plots show the first RHC planning step at $t = 0$ and the planned tug trajectories 24 hours ahead in time that collectively minimise the respective cost functions.

The tugs are limited to a maximum speed of 20 km/h, whereas the speeds of the oil tankers are randomly drawn from a uniform distribution on $[20, 30]$ km/h. Drift trajectories are perpendicular onto the patrol line, and crosspoints (red crosses) are generated from extrapolating the future positions of oil tankers and their resulting drift trajectories.

Note that compared with previous work, we have reduced the maximum speed of the tugs from 30 km/h to 20 km/h, thus making it more difficult for tugs to cover potential crosspoints. The speeds used above are in line with the literature (e.g., Det Norske Veritas (2009), Eide et al. (2007a)).

Another difference when compared with our previous work is the more realistic scenario of oil tankers leaving or entering the observation zone. For simplicity, we have implemented this feature such that whenever an oil tanker leaves the zone to the north or south, another oil tanker enters at the opposite end.

We have drawn random drift durations for each oil tanker from a uniform distribution on $[8, 9, \dots, 12]$, and set the alarm times used for f_3-f_6 to occur 3 hours after time of drift t_d . Hence, although we have shown crosspoints for the entire simulation interval, no crosspoints at time $t = 8-3 = 5$ or earlier will have an effect in the evaluation of cost functions f_3-f_6 . Likewise, without the alarm time, no crosspoints at time $t = 8$ or earlier will have an effect in the evaluation of cost functions f_1 and f_2 .

For f_5 and f_6 , we set the safe region $r = 50$ km, using a conservative value corresponding to half the maximum speed ($= 10$ km/h) times the number of hours until the first crosspoints can occur, namely 5.

Effect Of Alarm Time

Let us first examine the effect of our newly introduced alarm time t_a in cost functions f_3-f_6 , which means using the distances between crosspoints and the corresponding tug positions at the time of an alarm. Compared with f_1 and f_2 , it seems evident that including the alarm time causes the trajectories to better anticipate future crosspoints. For example, the planned trajectory for the bottom tug of f_3-f_6 turns north already around $t = 2$ to $t = 4$, whereas this turnaround does not occur until $t = 8$ or $t = 9$ for f_1 and f_2 .

Similarly, planned trajectories using f_3-f_6 clearly takes into account some future cost for the latter half of the simulation period, where many tugs turn the opposite direction of the tug trajectories planned using f_1 and f_2 .

In short, it appears that for any point in the tug trajectories, the algorithm asks itself "where should the tugs be some hours ahead in time when the first crosspoints can occur?" and directs the tugs accordingly.

Effect Of Squaring

The effect of squaring can be seen for f_2 vs. f_1 and for f_4 vs. f_3 . For f_2 (squared) vs. f_1 (non-squared), the planned trajectories for tugs are more likely to be positioned in-between crosspoints when using the squared cost function. For f_4 (squared) vs. f_3 (non-squared), another, related effect is visible for the bottom tug, which turns south to cover more of the southernmost cross points. Indeed, since squaring means punishing larger distances more, examination of several simulation scenarios not reproduced here shows that squaring leads to tugs spreading out more and covering larger areas.

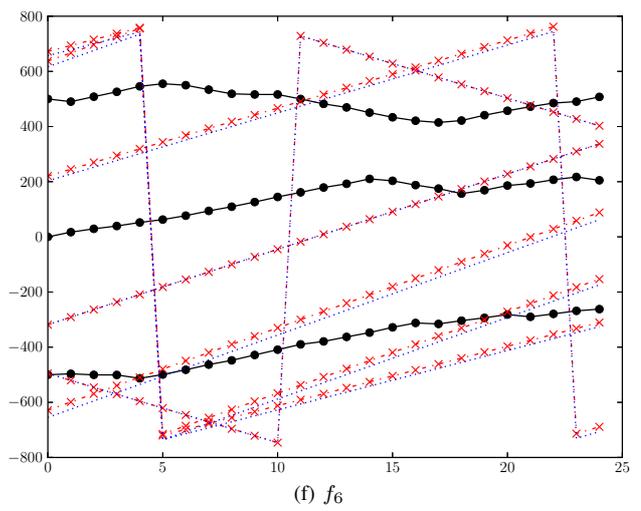
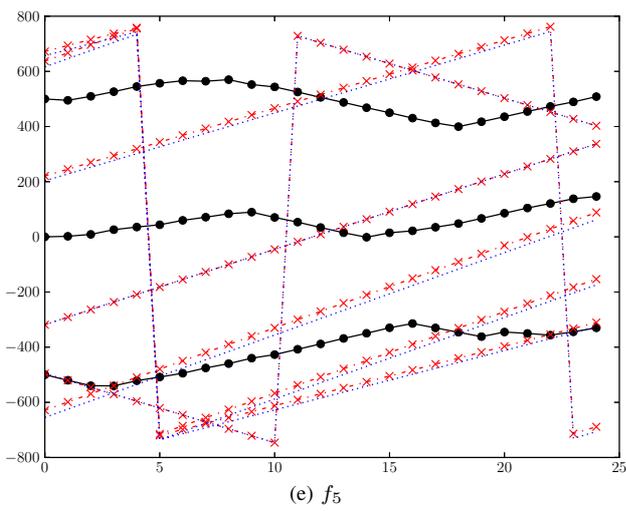
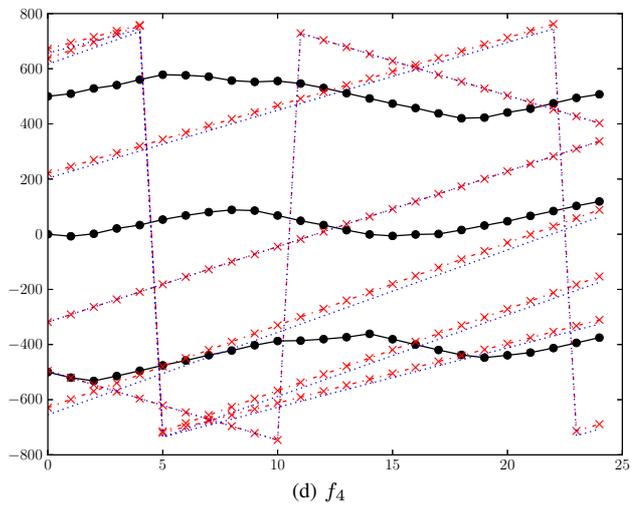
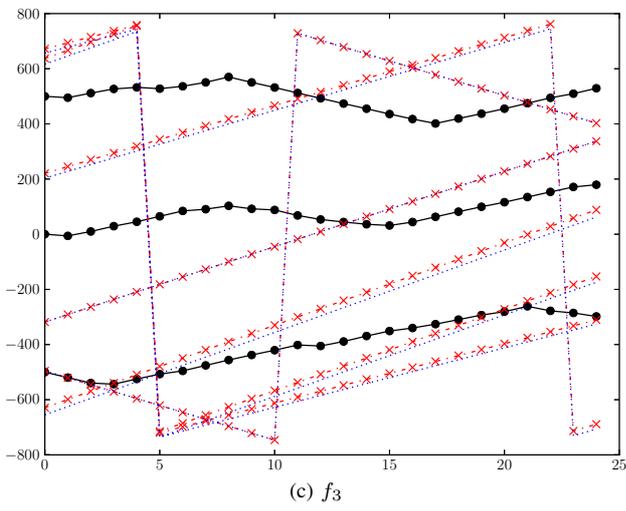
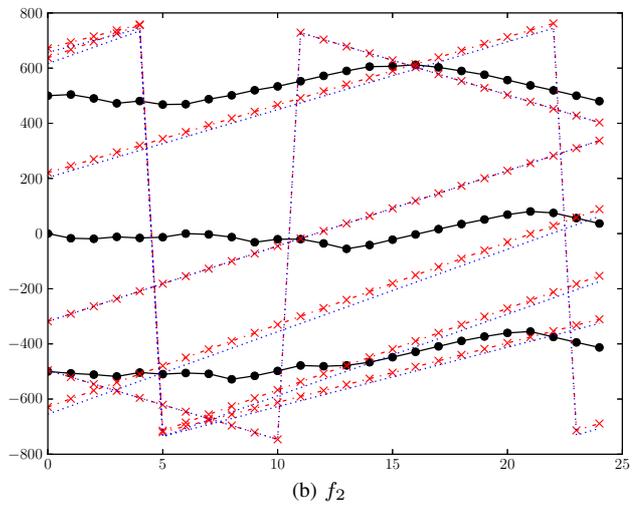
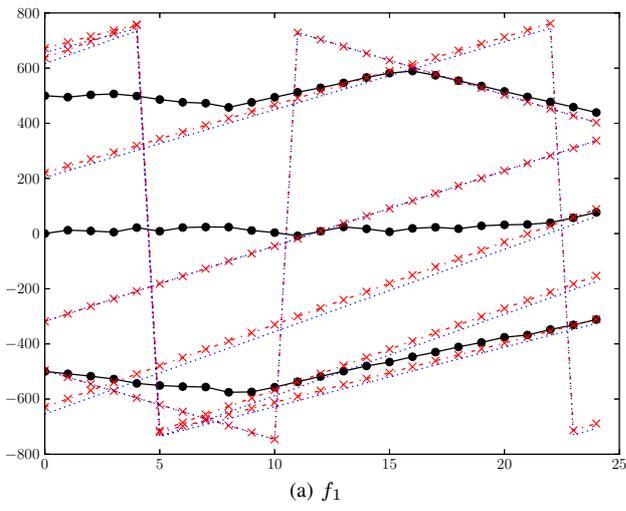


Figure 3: RHGA Planning Using Cost Functions f_1-f_6

Effect Of Safe Region r

The effect of employing a safe region $r = 50$ km is shown for f_5 and f_6 , which can be compared with f_3 that does not have a safe region. Comparing f_5 with f_3 , the two topmost tug trajectories are more or less the same, whereas using a safe region causes the bottom tug to travel south towards the end of the simulation scenario for f_5 . If we employ f_6 , which accumulates all non-reachable crosspoints and weights them all equally as a value of unity, the top tug trajectory is quite equal to that of f_3 , the middle tug trajectory seems a bit like a lagged version of the middle tug trajectory of f_3 and f_5 , and finally, the bottom tug trajectory of f_6 is quite stationary compared with f_3 and f_5 .

Objective Evaluation Of TFO Algorithms

Above, we suggested counting the number of unsalvageable oil tankers at every point in time as an objective evaluation method for comparing TFO algorithms. Here, we used this measure to compare each of the cost functions but there was little difference between the cost functions (not shown graphically). Examination of a number of simulation scenarios does seem to indicate that f_1 and f_2 perform worse than the others, but this needs a quantitative analysis to be confirmed.

Conclusions

In the above, we have made some qualitative interpretations of a particular simulation scenario with respect to our proposed cost functions. Similar qualitative interpretations have been made for a number of simulation scenarios not presented in this paper. Our study is largely work-in-progress and needs to be expanded upon to include quantitative analyses for comparing the properties, pros and cons, of different cost functions and TFO algorithms. Nevertheless, we think our observations are worthwhile, and in particular, we believe the simulation results add to our proposition that the original cost function f_1 (and f_2 , which is just a squared version of f_1) has a flaw by not including an alarm time and using the distances between tug positions at the time of an alarm and future positions of tankers after a DFA time $\hat{\Delta}_a$. Unfortunately, no strong conclusions can be made about our suggestion for an objective evaluation method for comparing cost functions just from the qualitative results presented here.

DISCUSSION

We have previously presented a formulation for a TFO problem and two algorithms, the RGHA and the RHMIPA, that solve it (Bye et al., 2010; Bye, 2012; Assimizele et al., 2013). Here, we identify a flaw in the cost function employed in the RHGA, namely the assumptions that tugs will be alarmed immediately when an oil tanker begins drifting, and that tugs will continue to execute their original plans even after receiving an alarm. Clearly, these assumptions are unrealistic. We point out how this flaw can

be corrected by means of defining a new, and shorter, drift-from-alarm (DFA) time, which is the estimated drift time from tugs receive an alarm until the drifting tanker crosses the patrol line. Incorporating this DFA time, we suggest several new cost functions that make comparisons between positions of tugs at the time of an alarm and the hypothetical future crosspoints of drifting tankers. This ensures that tugs abandon their planned original plans immediately upon receiving an alarm. The suggested cost functions can be used in a RHGA or in other TFO algorithms. We have tested the RHGA with a number of cost functions and simulation scenarios in our recently developed simulator framework. The simulation scenarios used in our previous work have been corrected and extended by allowing the realistic option of oil tankers both leaving and entering an observation zone, as well as lowering the maximum speed of patrol tugs to 20 km/h, thus making the problem even more complex to solve. Finally, we have suggested an objective evaluation measure for comparing various TFO algorithms and/or cost functions. More testing and analyses are needed in order to evaluate both the merit of the different cost functions and the suitability of the objective evaluation measure. The results are preliminary and a reflection of this paper as a report on work-in-progress but valuable nonetheless.

Simulator Framework

A thorough presentation of our new simulator framework requires a separate paper but we will cover the essential below. We chose the purely-functional programming language Haskell for our implementation, which means that functions in Haskell are pure, there is no global state, and no side effects. Code written in Haskell is therefore less error-prone and usually more concise, compact, and readable than imperative programming languages like C or Java. An additional advantages that “comes for free” with a functional language is a focus on *what* the programmer wants to achieve, rather than *how*, since functional program specifications can simply be executed directly rather than translated into imperative code.

A challenge, however, can be the use of pseudo-random number generators (PRNGs), which, by definition, are impure and require book-keeping of a system state. Nevertheless, using a functional language like Haskell provides a clear separation between pure and impure functionality, thus reducing this book-keeping to a minimum.

Whilst being strongly typed, and thus avoiding compile-time core dumps, Haskell uses polymorphism, which enhances the reusability of code. Many functions can therefore be written only once, because they accept input and output variables of many different types.

Haskell is also a good choice for parallel programming, which we believe is likely to be needed as the complexity of our simulator grows. Using pure parallelism guarantees deterministic processes and zero race conditions or deadlocks, however, non-pure concurrency related to PRNGs and other processes is also required.

Finally, it is worth mentioning that Haskell is a non-strict, lazy language, meaning that evaluation only happens on demand. This removes the need for the programmer to pre-allocate memory such as fixed-size arrays, and makes it easier to write modular programs, since functions can be passed freely to other functions, be returned as the result of a function, and stored in data structures.

Our choice of using Haskell for implementation makes our simulator very extendable and we are therefore confident that we will be able to perform several comprehensive and quantitative studies in the time to come.

Future Work

There are several directions the DRAMA research group wishes to pursue in the way forward. First of all, we need to perform an extensive simulation study based on what we have presented here. This will include simulating a large number of scenarios and examining if our proposed objective evaluation method can be used for comparing our proposed cost functions, and more generally, for comparing TFO algorithms.

Furthermore, our simulator needs to be extended to accommodate a large number of realistic simulation parameters and scenarios, including variable maximum speeds of tugs (the maximum speed of a tug constantly varies with wave height and sea roughness at the geographical location), tugs being temporarily unavailable (e.g., due to change of crew), realistic drift trajectories based on realistic models, fuel consumption and environmental impact, and 2D scenarios (e.g., oil tankers entering or leaving port represent high risk).

Moreover, we are already working on probabilistic models, for example for assigning risk weights to oil tankers depending on factors such as the geographical location or size and type of oil being carried; for generating continuous probability distributions of crosspoints; and for quantifying the probability of intercepting tankers in drift.

Finally, it would be of interest to include historical records of traffic data for both oil tankers and tugs in the simulator and determine the performance of the real-world tugs compared to various TFO algorithms. Such a comparison, if thorough and sound, can be used for analytical purposes and to provide support (or lack thereof) of using TFO algorithms as a decision-support tool at the VTS centres around the world.

ACKNOWLEDGEMENTS

The DRAMA research group is grateful for the support provided by Regionalt Forskningsfond Midt-Norge and the Research Council of Norway through the project *Dynamic Resource Allocation with Maritime Application* (DRAMA), grant no. ES504913.

REFERENCES

Assimizele, B., Oppen, J., and Bye, R. T. (2013). A sustainable model for optimal dynamic allocation of patrol tugs to oil tankers. In *Proceedings of the 27th European Conference on Modeling and Simulation*, pages 801–807.

- Breivik, Ø. and Allen, A. (2008). An operational search and rescue model for the Norwegian Sea and the North Sea. *Journal of Marine Systems*, 69(1–2):99–113.
- Breivik, Ø., Allen, A., Maisondieu, C., and Roth, J. (2011). Wind-induced drift of objects at sea: The leeway field method. *Applied Ocean Research*, 33(2):100–109.
- Bye, R. T. (2012). A receding horizon genetic algorithm for dynamic resource allocation: A case study on optimal positioning of tugs. *Series: Studies in Computational Intelligence*, 399:131–147. Springer-Verlag: Berlin Heidelberg.
- Bye, R. T., van Albada, S. B., and Yndestad, H. (2010). A receding horizon genetic algorithm for dynamic multi-target assignment and tracking: A case study on the optimal positioning of tug vessels along the northern Norwegian coast. In *Proceedings of the International Conference on Evolutionary Computation (ICEC 2010) — part of the International Joint Conference on Computational Intelligence (IJCCI 2010)*, pages 114–125.
- Det Norske Veritas (2009). Rapport Nr. 2009-1016. Revisjon Nr. 01. Tiltaksanalyse — Fartsgrenser for skip som opererer i norske farvann. Technical report, Sjøfartsdirektoratet.
- Eide, M. S., Endresen, Ø., Breivik, Ø., Brude, O. W., Ellingsen, I. H., Røang, K., Hauge, J., and Brett, P. O. (2007a). Prevention of oil spill from shipping by modelling of dynamic risk. *Marine Pollution Bulletin*, 54:1619–1633.
- Eide, M. S., Endresen, Ø., Brett, P. O., Ervik, J. L., and Røang, K. (2007b). Intelligent ship traffic monitoring for oil spill prevention: Risk based decision support building on AIS. *Marine Pollution Bulletin*, 54:145–148.
- Goodwin, G. C., Graebe, S. F., and Salgado, M. E. (2001). *Control System Design*. Prentice Hall, New Jersey.
- Hackett, B., Breivik, Ø., and Wettre, C. (2006). Forecasting the Drift of Objects and Substances in the Ocean. In *Ocean Weather Forecasting: An Integrated View of Oceanography*, pages 507–523. Springer.
- Havforskningsinstituttet (2010). Fisken og havet, særnummer 1a-2010: Det faglige grunnlaget for oppdateringen av forvaltningsplanen for Barentshavet og havområdene utenfor Lofoten. Technical report, Institute of Marine Research (Havforskningsinstituttet).
- Maciejowski, J. M. (2002). *Predictive Control with Constraints*. Prentice Hall, first edition.
- Rossiter, J. A. (2004). *Model-based Predictive Control*. CRC Press.

AUTHOR BIOGRAPHIES

ROBIN T. BYE² graduated from the University of New South Wales, Sydney with a BE (Hons 1), MEngSc, and a PhD, all in electrical engineering. Dr. Bye began working at the AAUC as a researcher in 2008 and has since 2010 been an associate professor in automation engineering at AAUC. His research interests include dynamic resource allocation, optimisation, computational intelligence, cybernetics, and human movement control systems.

HANS G. SCHAATHUN³ is cand.scient. 1999 and dr.scient. 2002 in coding theory from the University of Bergen and worked as lecturer and post.doc. there until 2006. He was lecturer and senior lecturer at the University of Surrey 2006–2010 specialising in multimedia security. Since 2011 he has been professor in computing at AAUC, focusing his research on modelling and simulation, machine learning, and software engineering.

²www.robinbye.com

³www.hg.schaathun.net

The value of integration in logistics

Claudia Archetti and M. Grazia Speranza
Department of Economics and Management
University of Brescia
I-25122, Brescia, Italy
Email: {archetti,speranza}@eco.unibs.it

KEYWORDS

Logistics; Integration; Optimization.

ABSTRACT

Many logistic problems arising in supply chain management, distribution and inventory management call for the integration of different components of the production/distribution system which have to be coordinated in such a way that a common objective, which can be the cost minimization or the revenue maximization, is optimized. Thus, in order to find the best management policy, one should be able to tackle the problem as a whole and to find an integrated policy that is aimed at optimizing the system behavior. However, known practices as well as the scientific literature have shown a major attitude in proposing strategies that are aimed at decomposing the system in parts and then proposing optimal policies for each single part. This clearly leads to a strategy that is far from being optimal for the global system. The aim of this work is to focus on the advantages that integrated policies can provide when used to handle production, inventory and distribution problems. We will present some cases that are mainly dealing with distribution problems and show the strategies proposed in the literature. We will also present a study on a distribution system where the inventory and the distribution costs have to be minimized.

INTRODUCTION

Logistics comprises all the activities related to the functioning of a production system or a supply chain in general. When talking about logistics many persons and professionals associate it with distribution and inventory management. However, logistics is much more than this. The *Council of Supply Chain Management Professionals* gives the following definition of *logistics*: 'Logistics management is that part of supply chain management that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers requirements'. Thus, the logistics management has an impact on all the activities of a supply chain. As these activities are linked, they need to be coordinated to guarantee a good performance of the supply chain, and the same holds for logistics. This creates the need to develop integrated management policies that tackle the system as a whole

and are pursued at optimizing the global performance of the system. Nothing new: it is well known that, in order to achieve the best overall performance, one has to optimize the system behavior as a whole. However, the major practices used by both professionals and scientists are based on decomposing the system and handling single parts independently. The main reason for this is related to the fact that integration is typically too difficult to achieve. Handling the global system as a whole often leads to a too complex optimization problem. However, on the other hand, decomposition leads to a worsening of the system performance which may be substantial. In fact, if on one side decomposition generates subproblems which are often much easier to handle than the integrated problem and thus for which an optimal strategy can be devised, on the other side, what is optimal for a subproblem rarely coincides with what is optimal for the integrated problem.

In the last years, the advances in technology, information systems, decision supporting tools and scientific research have favored the trend of considering larger and larger systems. The incentive in going in this direction comes from the economic advantages that may be achieved when improving the performance of the system through the development of an integrated policy.

Integration means not only finding the best overall policy, but also finding how to implement it in such a way that all the actors of the supply chain would accept it. In fact, a policy that optimizes the performance of the entire system may create advantages for some stakeholders and penalizes some other. Thus, it is crucial to define a policy to share the benefits among all the stakeholders.

In this paper we will focus on the first step of integration, i.e., the definition of integrated policies that optimize the overall system. We will not take into account the second step which deals with the definition of how the savings/revenues of this strategy should be shared among stakeholders. In particular, we will show some examples of savings achieved by integrated policies in problems arising in distribution. We will also present a computational study on an inventory routing problem which combines distribution and inventory management.

The paper is organized as follows. In the following section we present the class of problems on which we focus our study, that is the class of routing problems. We first describe the main setting of the problems and then present three examples of integration of routing deci-

sions with other strategic and/or operational decisions, namely, location, inventory management and loading. In Section II we present a computational study on the inventory routing problem where we show the advantages of the integrated policy with respect to different policies which optimize only a single component of the global objective. Finally, conclusions are drawn in Section III.

I. ROUTING PROBLEMS

Routing problems deal with the distribution of goods from one or several origins (suppliers) to one or several destinations (customers). Similar routing problems arise in collection problems where goods must be collected from origins and delivered to destinations. The distribution/collection operations are performed by means of a fleet of vehicles which are typically subject to a set of constraints such as capacity constraints, maximum duration of the route that a vehicle can carry out (that is related to the driver shift), starting and ending time of a route, time windows at the customers, etc. The objective is to find the vehicle routes, i.e., assign each customer to a vehicle and determine the sequence of visits of each vehicle, in such a way to minimize the operational cost which typically coincides with the total distance traveled by all vehicles. Another important component of the objective function is the total number of vehicles (and thus drivers) used to serve all customers.

Many variants of routing problems have been studied in the literature. For a survey, the reader is referred to [13]. Almost all variants of routing problems belong to the class of NP-complete problems, thus they are very complex problems for which the design of an optimal strategy is typically a hard task. This led the scientific community to follow two main research directions when dealing with routing problems. From a methodological point of view, the research has been mainly concentrated on the development of heuristic solution techniques, i.e., approaches that are aimed at providing good quality solutions without the guarantee of being optimal. This is due to the fact that the design of optimal solution methods is impractical in most routing applications. From an operational point of view, routing problems have been mainly studied as stand-alone problems, i.e., without the integration with other phases or activities of the production/distribution system they are part of. This choice is due to the fact that routing problems are already difficult in themselves, thus the effort has been focused on devising good solution techniques for the routing phase without enlarging the analysis to other operations. However, in the last years this trend has changed and the scientific literature is evolving towards the study of more integrated problems.

We now focus the study on three applications of the integration of routing decisions with other decisions taken at a strategic, tactical and operational level. At a strategic level we have the *Location Routing Problems* which combine routing and location decisions. At a

tactical level we study the *Inventory Routing Problem* combining routing and inventory management. Finally, at the operational level, we analyze the *Routing Problems with Loading Constraints* where the decision on how to serve the customers is combined with the one on how to load the goods on the vehicles.

A. Location Routing Problems

Following [12], *location routing* can be defined as ‘location planning with tour planning aspects taken into account’. Also, in [2], it is observed that ‘location/routing problems are essentially strategic decisions concerning ... facility location’. Thus, location routing problems are classified as strategic problems belonging to the research area of location theory and paying special attention to routing issues. They integrate the decisions on location of facilities, allocation of customers to facilities and definition of vehicle routes to serve such customers (see [11], [9]).

Distribution costs may play a crucial role in location decisions. In order to give an idea of how the distribution activities may influence the decision on where to locate facilities, we provide a simple example. Consider the problem where a company has three customers located in points A, B and C depicted in Figure 1. Suppose that the company has to decide where to locate a distribution center which will serve the three customers. If the company is planning to serve the three customers with a single vehicle serving the three customer together in a route, then the best place where to locate the distribution center is any point on the perimeter of triangle ABC. If instead each customer will be served independently from the others by a trip that goes directly from the distribution center to the customer and back, then the best place where to locate the distribution center is point O.

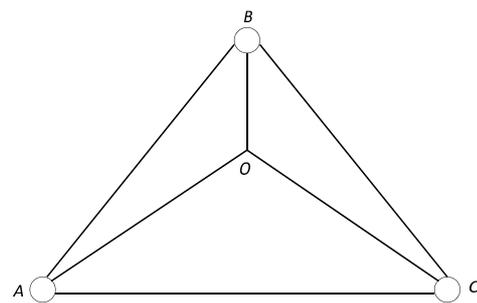


Fig. 1. Integrating location and distribution decisions

The literature on location routing problems has grown quite consistently in the last years as shown in the survey by Nagy and Salhi [12]. Location routing implies the adoption of an integrated view differing from classical location methodologies in the sense that also the routing aspects (and the related costs) are taken

into account. The interrelation between facility location and routing is well known as witnessed by a pioneering paper of the sixties by Maranzana [10]. However, both practitioners and researchers have ignored this interrelation for a long time solving location problems without considering the routing aspects. Following [12], this is mainly due to three reasons:

1. There are many practical applications where location problems do not have routing aspects.
2. Location and routing belong to two different decision levels: location decisions are strategic while routing is more tactical/operational. In fact, while routes can be changed quite frequently and without incurring in big reinvestments, it is much more difficult and expensive to change facility locations.
3. Location routing problems are more difficult and less tractable than classical location problems.

Despite the previous motivations, many practical application problems call for the integration of location and routing decisions. The scientific literature shows a number of papers dealing with real-life problems where a location routing model and solution technique has been applied (see [12] for more details). Most of them deal with the distribution of goods or parcels, but there are also applications in health, military and communications.

B. Inventory Routing Problems

Inventory routing problems deal with the integration of distribution planning and inventory management. They consider a supply chain where products have to be distributed from a supplier to a set of customers (typically retailers). The distribution plan has to take into account inventory constraints such as maximum inventory level at each customers (and possibly also at the supplier), no stockout or, on the contrary, possibility of backlogging, delivery frequency... The costs of the system typically include inventory costs at the supplier and at the customers and distribution costs. Two opposite policies may be thought of, and have been studied in the literature, when dealing with such a system:

- The *Retailer-Managed Inventory* (RMI) policy: in this policy, each customer (retailer) decides when he/she wants to be served and the quantity of product he/she wants to receive. The supplier plans the distribution on the basis on the decisions taken by all customers.
- The *Vendor-Managed Inventory* (VMI) policy: in this policy, the supplier monitors the inventory of each customer and determines its replenishment policy. At the same time, he/she defines the distribution plan. The RMI policy is a decentralized policy favoring the solution that maximizes the customers benefits. However, this policy may impose on the supplier constraints that increase its costs and thus, as a consequence, the costs of the product/service to its customers. The VMI policy can be seen as an integrated and centralized policy where the supplier is responsible for the management of the entire system. Obviously, the benefit generated from such a policy depends on the objective of

the supplier: he/she may take into account the global cost of the overall system or he/she may focus and give priority on some components of the system.

As pointed out in [5] and [6], the number of papers dealing with the VMI policy is growing. This is due to the advantages coming from the application of an integrated solution approach which is able to reduce the system costs.

In order to show the advantages of a VMI with respect to the RMI policy, consider the following simple example. A supply chain is formed by a supplier with a single warehouse which is used to serve three customers. The distribution plan has to be determined for the following 3 days. Each customer has a daily demand of 1 unit of product. The maximum inventory level at each customer is equal to 3 and no stockout is allowed. The distribution is performed through a fleet of homogeneous vehicles with a maximum capacity equal to 3 units. Each trip from the warehouse to a customer, as well as from a customer to another customer, has a cost of 1. The unitary inventory holding cost at all customers is 0.1. The inventory holding cost at the supplier is negligible. With the RMI policy, each customer is served every day with a delivery of 1 unit, as this is the policy that minimizes the inventory holding cost for each customer. This solution implies that a vehicle is used every day to deliver one unit to each customer, generating a daily distribution cost of 5 and, thus, a total distribution cost of 15 (over the three days). The inventory cost is equal to 0 as each customer receive 1 unit each day which is immediately consumed by the daily demand. Thus, the overall cost is 15. In the case of a VMI policy where the objective is to minimize the global system cost, the best solution is to send three vehicles from the warehouse to each customer on the first day, each delivering 3 unit to the customer. The distribution cost is 6. The inventory holding cost at each customer is equal to 0.3: on the second day the inventory level is equal to 2 units (3 units delivered on the first day minus 1 unit consumed at the first day) leading a holding cost of 0.2, while on the third day the inventory level is 1 and the corresponding holding cost is 0.1. Thus, the overall cost is 6.9.

The study of the potential benefits that may be achieved by an integrated policy in inventory routing problems dates back to the eighties with the pioneering paper by Bell et al. [3]. In [7] an extension of the problem is considered where production decisions are included.

In Section II, we will present a computational study to highlight the benefits that may be achieved from an integrated policy that minimizes the global cost of the system.

C. Routing Problems with Loading Constraints

Routing problems with loading constraints combine routing decisions, i.e., the assignment of customers to vehicles and the definition of the sequence of visits of customers for each vehicle, with loading decisions, i.e., how to load the goods on the vehicle. This problem is

encountered in many real-world transportation applications, especially when shippers deal with many items of different shapes and the loading aspect is not trivial. Examples include the distribution of furniture or mechanical components.

Routing problems with loading constraints may be classified as operational problems as they are related to the definition of the periodic distribution plan. They generalize the classical routing problems where a single product attribute is considered, namely, the weight. Thus, classical routing problems require that the total weight of the products loaded on each vehicle must be not greater than the vehicle capacity. This implies that the volume and the shape of the products do not have an influence when deciding how to assign products to vehicle. However, one may easily imagine that there exist a number of distribution problems where this assumption is not applicable.

The scientific literature classifies the routing problems with loading constraints in two main classes (see [8] and [14]):

- *Routing problems with two-dimensional loading constraints.* These problems arise in transportation applications dealing with items that cannot be stacked one on top of the other (because of their fragility or weight). This is the case of the transportation of refrigerators or pieces of catering equipment.
- *Routing problems with three-dimensional loading constraints.* In this case items can be superposed. An example is the transportation of furniture.

Loading items into two or three dimensional containers involve considerations not only on weight, shapes, fragility or possibility of superposing, but also on the order with which items are loaded on the vehicles. In fact, vehicle characteristics may influence the way items are loaded and unloaded. For example, for rear-loading vehicles, one has to first unload the items that are closer to the vehicle exit before being able to unload the ones that are on the back. These problems are called *routing problems with LIFO constraints*, meaning that items have to be unloaded with a reverse order with respect to the one used for the loading. This has a clear impact on the choice of the best way to serve the customers. As an example, consider the problem depicted in Figure 2. There are three customers and a single vehicle. The vehicle is located at a depot. Customers and depot locations are illustrated in Figure 2.c. Each customer requires a single item whose shape is depicted in Figure 2.a (the number on each item corresponds to the customer which requires it). The vehicle has a rear-loading container depicted in Figure 2.b. The three items completely occupy the container. The only way of feasibly loading all items in the vehicle is the one depicted in Figure 2.a or the opposite one, i.e., the one where item 2 is closer to the vehicle exit and item 1 is on the back. In any case, customer 3 has to be visited in between customer 1 and 2. If the LIFO constraints are ignored, an apparently better solution would be found, that consists in visiting customers 1 and 2 consecutively, as they are closer to each other with respect to customer 3.

This solution however would be impossible to implement because of the existing loading constraints.

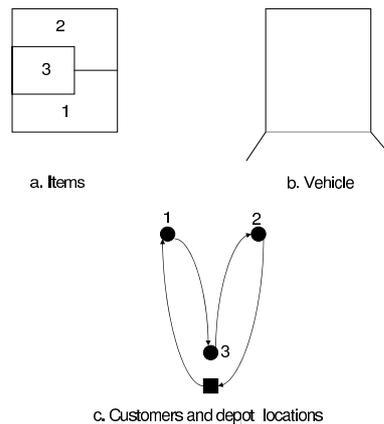


Fig. 2. A routing problem with LIFO constraints

II. COMPUTATIONAL STUDY

In this section we present a computational study focused on the analysis of the benefits of an integrated policy applied to the inventory routing problem. We consider a supply chain where a supplier, denoted as node 0, has to deliver a product to a set \mathcal{M} of geographically dispersed customers with $|\mathcal{M}| = n$. A planning horizon \mathcal{T} is defined and time is discretized in periods, say days. We denote as H the number of periods in the planning horizon. Each customer s faces a daily demand and defines a maximum inventory level U_s . We denote as r_{st} the demand of customer s at day t . No stockout is allowed at the customers. The product is made available at the supplier at a daily rate r_{0t} and no maximum inventory level is defined for the supplier. The distribution of the product to the customers is made through a single vehicle with a given maximum capacity C . Thus, the total quantity loaded on the vehicle in each day must not exceed the vehicle capacity. The vehicle is located at the supplier. We denote as c_{ij} the cost of going from node i (either a customer or the supplier) to node j and the unitary inventory holding cost for customer s is denoted as h_s (similarly, h_0 denotes the unitary inventory holding cost at the supplier). The system cost comprises the following three terms:

1. the distribution cost: this is given by the cost of the distance traveled by the vehicle to transport the product from the supplier to the customers in each day of the planning horizon;
2. the inventory cost at the customers: each customer faces a unitary daily inventory holding cost for the product held in inventory in each day of the planning horizon;
3. the inventory cost at the supplier: similarly to the case of the customers, the supplier faces a unitary daily inventory holding cost for the product held in inventory in each day of the planning horizon.

Our study will compare the following policies:

- VMI: the objective is to minimize the global system cost composed by all three terms previously defined.
- RMI: a hierarchical objective is established. The main objective is the minimization of the inventory cost at the customers. The secondary objective is the minimization of the sum of the distribution costs and the inventory costs at the supplier. This corresponds to a sequential optimization where first the customers optimize their own decisions and then the supplier optimizes its costs, taking the customers decisions as constraints.

A mathematical formulation of the VMI policy is presented in [1]. We report it here for the sake of completeness. The formulation makes use of the following variables:

1. B_t : inventory level at the supplier at day t ;
2. I_{st} : inventory level of customer s at day t ;
3. z_{it} : binary variable indicating whether node i (either a customer or the supplier) is visited at day t ;
4. x_{st} : quantity delivered to customer s at day t ;
5. y_{ij}^t : integer (binary) variable indicating the number of times the vehicle travels from node i (either a customer or the supplier) to node j at day t .

$$\min \quad \sum_{t \in \mathcal{T}'} h_0 B_t + \sum_{s \in \mathcal{M}} \sum_{t \in \mathcal{T}'} h_s I_{st} + \sum_{i \in \mathcal{M}'} \sum_{j \in \mathcal{M}', j < i} \sum_{t \in \mathcal{T}} c_{ij} y_{ij}^t \quad (1)$$

$$\text{s.t.} \quad B_t = B_{t-1} + r_{0t-1} - \sum_{s \in \mathcal{M}} x_{st-1} \quad t \in \mathcal{T}' \quad (2)$$

$$B_t \geq \sum_{s \in \mathcal{M}} x_{st} \quad t \in \mathcal{T} \quad (3)$$

$$I_{st} = I_{st-1} + x_{st-1} - r_{st-1} \quad s \in \mathcal{M} \quad t \in \mathcal{T}' \quad (4)$$

$$I_{st} \geq 0 \quad s \in \mathcal{M} \quad t \in \mathcal{T}' \quad (5)$$

$$x_{st} \leq U_s - I_{st} \quad s \in \mathcal{M} \quad t \in \mathcal{T} \quad (6)$$

$$x_{st} \leq U_s z_{st} \quad s \in \mathcal{M} \quad t \in \mathcal{T} \quad (7)$$

$$\sum_{s \in \mathcal{M}} x_{st} \leq C \quad t \in \mathcal{T} \quad (8)$$

$$\sum_{s \in \mathcal{M}} x_{st} \leq C z_{0t} \quad t \in \mathcal{T} \quad (9)$$

$$\sum_{j \in \mathcal{M}', j < i} y_{ij}^t + \sum_{j \in \mathcal{M}', j > i} y_{ji}^t = 2z_{it} \quad i \in \mathcal{M}' \quad t \in \mathcal{T} \quad (10)$$

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}, j < i} y_{ij}^t \leq \sum_{i \in \mathcal{S}} z_{it} - z_{kt} \quad \mathcal{S} \subseteq \mathcal{M} \quad t \in \mathcal{T} \quad (11)$$

$$x_{st} \geq 0 \quad s \in \mathcal{M} \quad t \in \mathcal{T} \quad (12)$$

$$y_{ij}^t \in \{0, 1\} \quad i \in \mathcal{M} \quad j \in \mathcal{M}, j < i \quad t \in \mathcal{T} \quad (13)$$

$$y_{i0}^t \in \{0, 1, 2\} \quad i \in \mathcal{M} \quad t \in \mathcal{T} \quad (14)$$

$$z_{it} \in \{0, 1\} \quad i \in \mathcal{M}' \quad t \in \mathcal{T}. \quad (15)$$

Note that $\mathcal{T}' = \mathcal{T} \cup \{H+1\}$. We consider day $H+1$ to take into account the inventory costs at the end of the planning horizon. The objective function (1) minimizes the total cost. Constraints (2)–(5) define the inventory level at the supplier and at the customers and impose to have no stock-out. (6) and (7) define the maximum inventory level at the customers. Constraints (8) and (9) are vehicle capacity constraints. (10) and (11) are routing constraints. In particular, (10) establish that if a node is visited at day t , then the vehicle has to enter

and to exit from the node. Constraints (11) are subtour elimination constraints. Finally, (12)–(15) are variable definitions. Note that formulation (1)–(15) corresponds to a Mixed Integer Linear Program (MILP) and can thus be solved to optimality using a standard solution method for MILPs.

For the computational study, we use the branch-and-cut algorithm proposed in [1] and change the objective function according to the policy. This allows us to compare the optimal solution of each policy. The branch-and-cut algorithm proposed in [1] is a branch-and-bound algorithm where the subtour elimination constraints (11), which are exponential in number, are inserted only when violated, as in standard branch-and-cut algorithms for routing problems.

For the VMI policy the objective function corresponds to (1). For the RMI policy, the objective function is the following:

$$\sum_{t \in \mathcal{T}'} h_0 B_t + M \sum_{s \in \mathcal{M}} \sum_{t \in \mathcal{T}'} h_s I_{st} + \sum_{i \in \mathcal{M}'} \sum_{j \in \mathcal{M}', j < i} \sum_{t \in \mathcal{T}} c_{ij} y_{ij}^t, \quad (16)$$

where M is a large value. Thus, the optimization will first optimize the dominant part of the objective function, that is the inventory costs at the customers, and then, given that part of the solution, the costs of the supplier.

A similar study was reported in [4] where different policies were compared. The differences with respect to the current study are two-fold. Firstly, in [4] the order-up-to level policy is studied, which is a distribution policy imposing that, each time a customer is served, the quantity delivered is such that the maximum inventory level is reached. Secondly, the comparison of [4] is performed through the use of a heuristic algorithm, while in our study we compare the optimal solutions. Thus, the differences we report are more reliable in the sense that they do not depend on the quality of the solution provided by a heuristic algorithm.

Tests are performed on a subset of the instances proposed in [1] with the following characteristics:

- planning horizon: 3 and 6 days;
- number of customers: 30;
- inventory costs at the customers: randomly generated in the intervals $[0.01; 0.05]$ and $[0.1; 0.5]$;
- inventory costs at the supplier: equal to 0.03 when the inventory cost at the customers is generated in the interval $[0.01; 0.05]$ and equal to 0.3 when the inventory cost at the customers is generated in the interval $[0.1; 0.5]$;
- distribution costs: equal to the Euclidean distances between the location of the customers and the supplier, which are defined through Euclidean coordinates.

For each of the previous characteristics, 5 different instances were created for a total of 20 instances. For more details about the instances, the reader is referred to [1].

Results are shown in Figures 3-6. In Figure 3 the average total cost is reported. The three terms of the total cost are represented in the following three figures:

the inventory cost at the customers (Figure 4), the inventory cost at the supplier (Figure 5) and the distribution cost (Figure 6). Each figure reports, for each policy, the average results in the following settings:

- planning horizon of 3 days ($H = 3$);
- planning horizon of 6 days ($H = 6$);
- inventory costs at the customers randomly generated in the interval $[0.01;0.05]$ (Low inv. cost);
- inventory costs at the customers randomly generated in the interval $[0.1;0.5]$ (High inv. cost);
- entire test bed (Total).

Focusing on Figure 3, we can see that the RMI policy leads to solutions with a global cost that is much higher than the one obtained through the VMI policy. This is an indicator of the benefits that may be achieved by an integrated policy when considering the performance of the global system. In particular, the average increase of the total cost over all instances is 34.21% while the maximum increase is 65.61%. The highest difference is obtained in the case of a long planning horizon ($H = 6$) and low inventory cost. The longer the planning horizon is, the larger is the deterioration of the global cost due to the fact that the decisions taken by the customers about when to be served and how much to receive have a stronger impact on the distribution cost and the inventory cost at the supplier. Moreover, when the inventory cost is low, the distribution cost has a higher impact on the global cost. Thus, the RMI policy tends to produce worse solutions in terms of global cost.

Figure 4 shows how the VMI policy increases the inventory cost at the customers. This happens in particular when the planning horizon is long and when the inventory cost is high. The average increase of the inventory cost at the customers due to the VMI policy is 104.38% while the maximum increase is 173.5%. This clearly indicates that an integrated policy must be necessarily accompanied with a strategy defining how the global benefits must be shared among the stakeholders.

The difference in terms of inventory cost at the supplier is not substantial as shown in Figure 5. The average increase of the inventory cost at the supplier due to the RMI policy is 12.06% while the maximum increase is 17.32%. If we consider the distribution cost (Figure 6), the differences are remarkable: the average increase related to the RMI policy is 64.36% while the maximum increase is 83.09%. This shows that the distribution cost has a high impact on the global cost in the test instances we have studied.

To summarize, our computational study has shown that integration may indeed generate relevant benefits in terms of system optimization. However, these benefits must be properly shared among all stakeholders of the system. In fact, the reduction of the global system cost may lead to an increase of some specific components of the total cost that in turn has a negative impact on some of the stakeholders of the system.

III. CONCLUSIONS

Globalization and competition, combined with the development of information and communication tech-

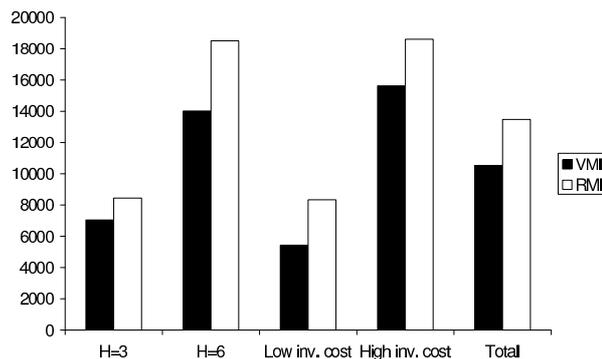


Fig. 3. Comparison of policies with respect to the total cost

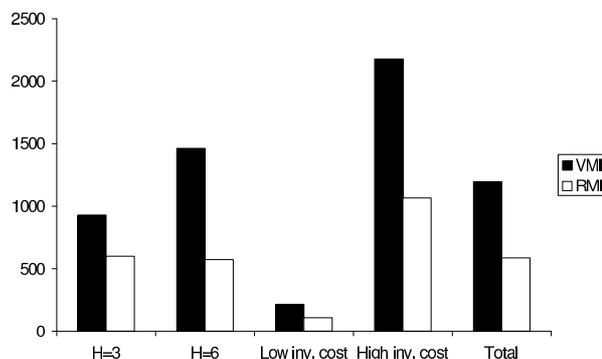


Fig. 4. Comparison of policies with respect to the inventory costs at the customers

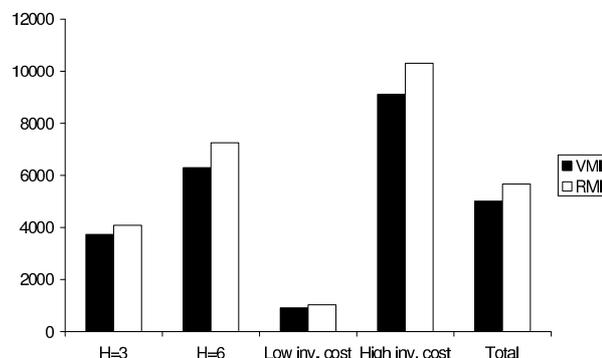


Fig. 5. Comparison of policies with respect to the inventory cost at the supplier

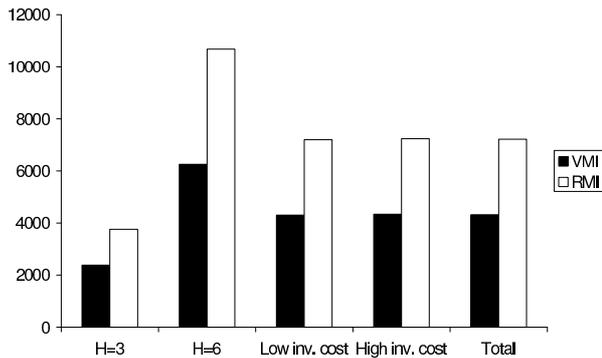


Fig. 6. Comparison of policies with respect to the distribution cost

nology, have pushed supply chains towards a global optimization process.

In the scientific community there is a general still growing trend towards modeling and solving optimization problems in logistics that jointly consider problems that were traditionally treated independently from each other. In this paper we have shown, also through a computational study, that this leads to better solutions. In particular, the computational study is focused on the inventory routing problem which combines distribution operations with inventory management. We have shown that an integrated policy which optimizes the system cost can achieve great benefits when compared to a decentralized policy where each stakeholder optimizes its own costs. These benefits generate an improved performance of the global system which can have positive returns on all the actors in case a policy is thought to properly share these benefits.

REFERENCES

- [1] C. Archetti, L. Bertazzi, G. Laporte, M.G. Speranza (2007). A branch-and-cut algorithm for a vendor managed inventory routing problem. *Transportation Science* 41, 382-391.
- [2] A. Balakrishnan, J.E. Ward, R.T. Wong (1987). Integrated facility location and vehicle routing models: Recent work and future prospects. *American Journal of Mathematical and Management Sciences* 7, 35-61.
- [3] W.J. Bell, L.M. Dalberto, M.L. Fisher, A.J. Greenfield, R. Jaikumar, P. Kedia, R.G. Mack, P.J. Prutzman (1983). Improving the distribution of industrial gases with an online computerized routing and scheduling optimizer. *Interfaces* 13, 4-23.
- [4] L. Bertazzi, G. Paletta, M.G. Speranza (2002). Deterministic order-up-to-evel policies in an inventory routing problem. *Transportation Science* 36, 119-132.
- [5] L. Bertazzi, M. Savelsbergh, M.G. Speranza (2008). Inventory routing. In: B. Golden, S. Raghavan, E. Wasil (Eds.), *The Vehicle Routing Problem: Latest Advances and New Challenges*. Springer, New York, 49-72.
- [6] L. Bertazzi, M.G. Speranza (2012). Inventory routing problems: An introduction. *EURO Journal on Transportation and Logistics* 1, 307-326.
- [7] P. Chandra, M.L. Fisher (1994). Coordination of production and distribution planning. *European Journal of Operational Research* 72, 503-517.
- [8] M. Iori, S. Martello (2010). Routing problems with loading constraints. *TOP* 18, 4-27.
- [9] G. Laporte (1988). Location-routing problems. In: B.L. Golden, A.A. Assad (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, 163-198.

- [10] F.E. Maranzana (1964). On the location of supply points to minimise transport costs. *Operational Research Quarterly* 15, 261-270.
- [11] H. Min, V. Jayaraman, R. Srivastava (1998). Combined location-routing problems: A synthesis and future research directions. *European Journal of Operational Research* 108, 1-15.
- [12] G. Nagy, S. Salhi (2007). Location-routing: Issues, models and methods. *European Journal of Operational Research* 177, 649-672.
- [13] P. Toth, D. Vigo (2006). *The Vehicle Routing Problem*. SIAM, Philadelphia.
- [14] F. Wang, Y. Tao, N. Shi (2009). A survey on vehicle routing problem with loading constraints. *International Joint Conference on Computational Sciences and Optimization* 2, 602-606.

CLAUDIA ARCHETTI is Assistant Professor of Operational Research at the University of Brescia, Italy. She has got a PhD in 'Computational Methods for Economic and Financial Decisions and Forecasting'. Her research interests are related to combinatorial optimization, routing problems, supply chain management. She is Associate Editor of Networks. She is a member of the Operational Research Group of the University of Brescia: <http://or-brescia.unibs.it/>.

M. GRAZIA SPERANZA is Full Professor of Operational Research at the University of Brescia, Italy. She was President of EURO, the Association of European Operational Research Societies, and Vice-President of IFORS, the International Federation of Operations Research Societies. She is currently President of the Transportation Science and Logistic society of INFORMS, the Institute for Operations Research and Management Science. Her research interests include combinatorial optimization, worst-case analysis, routing problems, supply chain management, optimization in finance. She is the leader of the Operational Research Group of the University of Brescia: <http://or-brescia.unibs.it/>.

Finance, Economics and Social Science

PETRI NETS AS TOOLS FOR POLICY ANALYSIS: THE EXAMPLE OF SMOKING BANS IN PUBLIC PLACES

Georg P. Mueller
Faculty of Economics and Social Science
University of Fribourg
Blvd de Pérolles 90
CH - 1700 Fribourg, Switzerland
E-mail: Georg.Mueller@Unifr.ch

KEYWORDS

Petri nets, policy analysis, smoking bans, public health, semi-quantitative methods.

ABSTRACT

Petri nets were originally developed in order to describe concurrent processes in computers and other automata. This paper identifies the features of Petri nets, which are essential for doing semi-quantitative analyses of public policies. The toolkit resulting from this methodological investigation is subsequently used for an exemplary analysis of smoking-bans and -restrictions in public places. At the end, the article presents empirical evidence, which seems to corroborate some inferences deduced from the model.

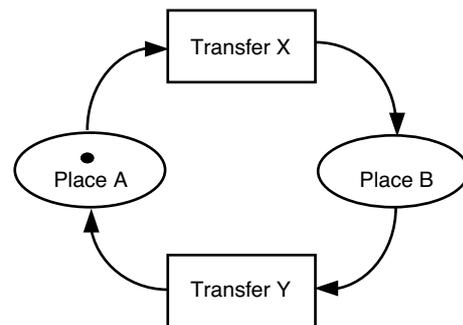
INTRODUCTION

This article proposes to use *Petri nets* (Van der Aalst and Stahl 2011: chap. 3; Wang 1998: chap. 2) for modelling *public policies* (Dunn 2004: chap. 2; Parsons 1997: chap. 3; Rossi and Freeman 1993: chap. 1). At first, the two topics seem to be rather disparate. *Public policy analysis* deals with political processes like the introduction of smoking bans or other measures for the promotion of public health. Among others, public policy analysis asks about

- the political and social *processes* required for explaining the existence or non-existence of certain public policies in different national contexts;
- the *feasibility* of planned new policy-goals;
- the existence of *concurrent harmful processes*, which may unintentionally be triggered by new policies.

Petri nets – on the other hand – have been developed for describing technical systems like computers or vending automata. They use for this purpose a special graphical language, which allows to model the transfers of so called tokens or objects from one place in the system to another. Fig. 1 depicts such an exemplary network, where tokens are being moved between two places A and B by cyclical transfers X and Y.

In spite of these differences, Petri nets reveal to have functionalities, which make them much more useful for policy analysis than originally thought (Heitsch et al. 2001; Koehler et al. 2001a; Koehler et al. 2001b). Like public policy analysis, Petri nets are *process-oriented*. They are especially made for describing *concurrent* processes, which are also a major concern of public policy analysis: as mentioned before, public policy analysis has e.g. an interest in the unintended consequences of new policies, which often run in parallel with their implementation. Similarly, it often occurs that new policies can only be put into effect, if concurrent secular processes like the availability of material resources or popular support have reached a certain stage. Besides, a branch of Petri net analysis focuses on the *reachability* of goals or places (Wang 1998: chap. 2.4.1), which promises to be also of use for feasibility studies in political planning. Finally, Petri nets can be used for modelling networks with *limited transfer capacities*, which is obviously also an important topic for the implementation of new public policies in state agencies.



Legend: ●: Token, being cyclically transferred between places A and B.

Fig. 1: A small exemplary Petri net.

All these communalities justify to consider Petri nets as potential tools for public policy analyses. Hence, this paper will first identify those *features* of Petri net analysis that are especially useful for modelling and investigating public policies. These features will then be used for analysing an exemplary case: the European anti-smoking policies, which have been put into effect

in the last 25 years, in order to prevent passive smoking in public places (see Goel and Nelson 2008).

PETRI NET MODELS FOR PUBLIC POLICY ANALYSIS

The Main Elements for Constructing Social Models

As mentioned before, Petri nets have been developed for describing and analysing technical systems. Consequently, when being used for public policy analysis, at least some of their features require an appropriate adaptation to the *specificities of social sciences*, which will be given in the sections that follow.

One of the central concepts of Petri nets are *tokens*, that can be „anything“ and may even have different qualities in the same model. The *main* category of tokens for policy analysis are *social problems*, which are transferred from one institution or office of the state to the next in order to be transformed into solutions. At the higher level of abstraction, which refers to *policy making*, tokens are *requests* from the public and the final outcome of the mentioned successive transformations are new *laws* or new administrative *procedures*. At the lower level of abstraction, mainly referring to *policy application*, social problems are *cases* like e.g. individuals, which lose by successive transformations by state agencies, such as prisons or unemployment offices, their problematic facets. Obviously, there are also *subsidiary* „technical“ tokens, which are required to model concurrent social processes, like e.g. the availability of power, money, etc. They often have system specific numerical values and thus create a *coloured* Petri net (see e.g. Jensen 1992: chap. 2.1).

Another central concept of Petri nets are *places*. For policy analysis we will consider them as *political institutions* or *state agencies*. Contrary to the situation in technical Petri net analysis, the incumbents of places have intentions and plans, which sometimes are even in conflict with the plans represented by other places. Consequently five types of particular graphic network elements are especially important for policy analyses:

- Priorities* of alternative transfers of tokens, indicated by *i*, *ii*, ... , etc. In Fig. 2 e.g., the incumbent of place C has a preference for transfer Z1 over transfer Z2.
- Inhibitions* of transfers. In the exemplary Fig. 2, the token at place A is able to inhibit the transfer W.
- Triggers* of transfers. In Fig. 2 e.g., the token at place A is also able to trigger the transfer X.
- Tokens with *time dependent* numerical values, which have a *white* filling. In the exemplary Fig. 2 there is such a token at place A, where it may be used for describing a secular development of a political system.
- Capacities* of places for the simultaneous *treatment* of social problems. In Petri net diagrams they are indicated by small bold figures **1**, **2**, ... assigned to

places like B or C in Fig. 2, which consequently can host the number of tokens corresponding to these figures. If there is no explicit information about the treatment-capacity of an agency or political institution, it is by default equal to 1.

With the mentioned special elements (a) to (e) and the standard symbols for tokens, places, and transfers, it should be possible to construct a qualitative model of nearly any policy process. Useful for this kind qualitative modelling are not only the theoretical literature but also abstractions from typical cases. Hence, Petri net modelling is relatively close to established *qualitative research* techniques like *grounded theory* (Strauss and Corbin 1998) or *qualitative comparative analysis* (QCA) (Ragin 2007). Petri net modelling for policy analysis may however also have *quantitative* facets: transfers may e.g. be triggered, if the time-dependent value *v* of a white-coloured token falls below a certain threshold *e* (see Fig. 2, place A). In the latter case, classical statistical research techniques like logistic regression (see e.g. Aldrich and Nelson 1992) may become important for *estimating* the values of these thresholds.

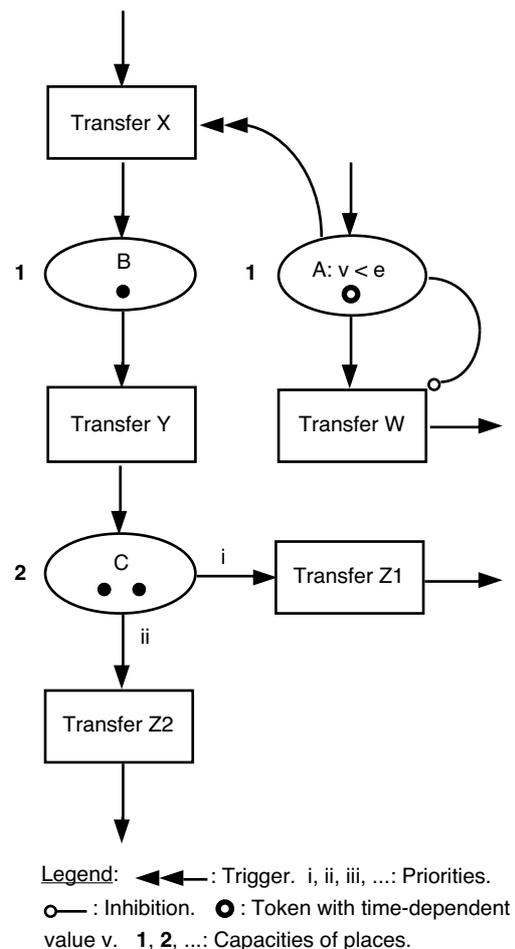


Fig. 2: Essential elements of Petri nets for public policy analysis.

Model Exploration and Model Testing

Petri nets can be „fed“ with two types of tokens: *Empirical* tokens with observational data from real politics and *experimental* tokens with fictitious values and properties. Empirical tokens are generally used for *model testing*: here, the focus is on the correspondence between the outcome of the model and the related empirical observations, often evaluated by conventional statistical test procedures (see e.g. Newcomer 1994). To the contrary, *experimental* tokens are primarily for model *exploration*, which means in this context ex-ante and ex-post evaluation. Experimental tokens allow to assess ex-ante the *feasibility* of new policies as well as their *unintended negative consequences*. By means of experimental tokens, it is also possible to *criticise* past policies, e.g. if an ex-post analysis reveals that there would have been better alternatives.

Model exploration with experimental tokens is traditionally done „by hand“ (see e.g. Reisig and Rozenberg 1998). In computer science there is a long tradition of mathematical reasoning in order to proof e.g. that certain places within a Petri net are reachable or absorbing deadlocks. This intellectual tradition is certainly also useful for the analysis of Petri net models of public policies. With only a small mathematical effort it is e.g. possible to evaluate the reachability of goals and thus the feasibility of the corresponding policy.

However, if empirical tokens are used to test a model with a great number of real cases, mathematical „handwork“ can be tedious. In this situation it is advisable to *simulate* the dynamics of a model. The database of Heitmann (2013) mentions for this purpose a great number of specialised software products. Alternatively, the much better known EXCEL may also be used for such investigations. In this case, the *lines* of the EXCEL spreadsheet represent subsequent time-points, whereas the *columns* are Petri net places with codes 0, 1, 2, ... , representing the number of tokens, by which they are occupied at a given time-point. Once the essential parts of a Petri net model are programmed, its long-term dynamics can be explored by iteratively repeating this program-nucleus with EXCEL's copy-paste function. Needless to say that such an EXCEL program is also very useful for trial-and-error optimisation of policies.

AN EXEMPLARY APPLICATION: SMOKING BANS

The Explanandum

For many years, tobacco taxes have been the main instrument for controlling the consequences of excessive smoking behaviour (Goel and Nelson 2008: chap. 2). Since this policy was obviously not very efficient in protecting non-smokers against nicotine abuse by others, many governments introduced after 1990 smoking restrictions and smoking bans in public places like restaurants, railway stations, etc. (Goel and Nelson 2008:

chap. 7; WHO 2007: part 2). As analysed by the author in an earlier publication (Mueller 2013), preventive measures against passive smoking generally started with smoking restrictions, e.g. by the introduction of smoke-free zones. A few years later, such policies were often replaced by total smoking bans in public places. In the mentioned publication the author was able to show that the schedule of this process depends on the *share of the smokers* in the total adult population. In democratic countries, smokers have voting power in parliamentary elections, which has to be considered by the legislators. Consequently, the higher the share of smokers in the adult population, the slower the stepwise process from, *unrestricted smoking to smoking restrictions* and finally to *smoking bans*. In what follows, we will describe this process by a Petri net model, among others in order to explain the different national histories of smoking-bans and -restrictions.

The Model

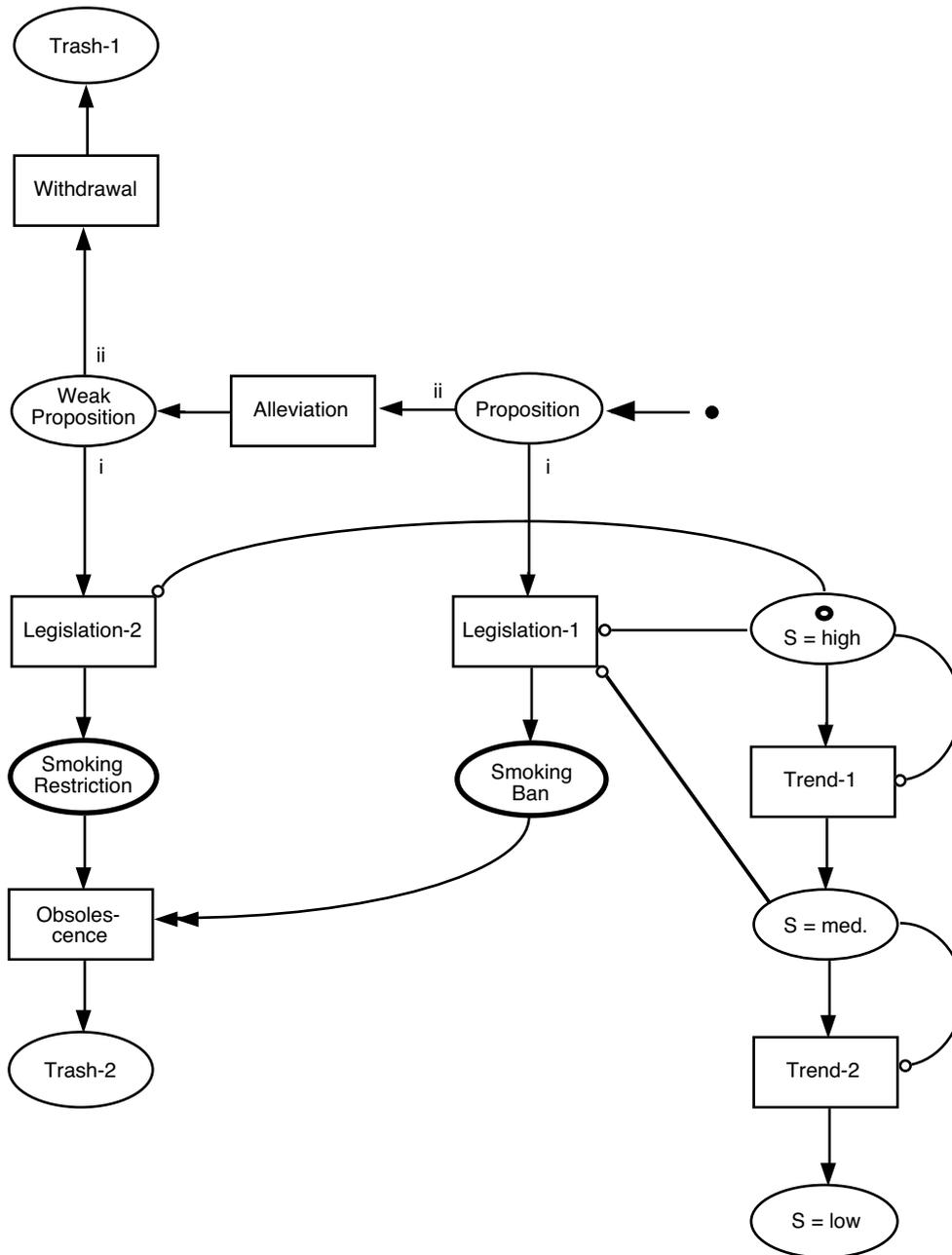
The model, which we propose in Fig. 3 for explaining the introduction of smoking-bans and -restrictions is driven by two *concurrent* processes. The first refers to the processing of *black* tokens representing anti-smoking proposals by the political system. The second process describes the „demography“ of smokers and non-smokers by means of a *white* token, which represents the *smoking rate S*, i.e. the percentage of the adult national population, which is regularly smoking. This process is considered to be rather autonomous, i.e. not directly influenced by smoking bans, but rather by cultural factors like e.g. the spread of healthier life-styles.

At the start t_0 of the process, the *white token* is always in the top-category „S=high“. Thus we assume that smoking rates were in the 1980-ies anywhere in Europe relatively high and tended to drop only in the subsequent years to the currently observed lower levels. The second part of this assumption means that the white token systematically decreases its value over time, until it falls below a threshold e' , which separates the high smoking rates from the others. Consequently, „Trend-1“ in Fig. 3 is no more inhibited and the token is transferred to the correct medium category of smoking „S=med.“. By a further decrease, the smoking rate finally falls below a second threshold e such that the inhibition of „Trend-2“ is also lifted and the token drops into the final category „S=low“.

As Fig. 3 demonstrates, the afore-mentioned dynamics of the share of the smokers has immediate consequences for the processing of new anti-smoking propositions, depicted as *black tokens*. As long as the smoking rate is high and „S=high“ is occupied by a white token, both transfers „Legislation-1“ and „Legislation-2“ are blocked by a strong smoking-lobby and propositions cannot be transformed in a legally binding „Smoking Ban“ or „Smoking Restriction“. Consequently, new propositions will be withdrawn and thus go to „Trash-1“. As soon as the smoking rate drops to

the medium level „S=med.“, the inhibition of „Legislation-2“ is lifted and a „Smoking Restriction“ becomes possible. Thus, after an „Alleviation“ of the initial proposal that makes it politically more acceptable to the shrinking share of smokers, the policy maker follows its first priority i and realises a „Smoking Restriction“. If at the end of long secular process the white token is in category „S=low“, the electoral power of the smokers is so weak that not only „Smoking Restriction“ but also „Smoking Ban“ are viable policies, since in this situa-

tion all inhibitions of „Legislation-1“ and „Legislation-2“ are lifted. It is assumed that the policy maker has an interest in strong anti-smoking laws and consequently follows its priority i: Instead of proceeding to the „Alleviation“ of new proposals, they are transferred to „Legislation-1“ and put into effect as „Smoking Bans“. As indicated in Fig. 3, this triggers the „Obsolescence“ of older smoking restrictions, which are consequently suspended.



Legend: S = Smoking rate. S=high: $S \geq e'$. S=med.: $e' > S \geq e$. S=low: $S < e$. Other symbols: See text and Fig. 2.

Fig. 3: A Petri net model of smoking-bans and -restrictions:
The situation at time t_0 .

Some Inferences from the Model

In principle, it would be possible to translate Fig. 3 into an EXCEL program that simulates the processing of artificial and real country tokens. However, inferences extracted this way are often not very transparent, mainly due to the absence of rationales. Hence, we preferred to do the simulation „by hand“ and extracted this way four major results:

- a) The way from „free“, unrestricted smoking to smoking restrictions and smoking bans is *unidirectional* with no possibility to return to an earlier stage in this sequence. This is a consequence of the unidirectional decrease from high to low smoking rates as well as of the preferences of the policy maker for more restrictive anti-smoking measures.
- b) It is very *unlikely* that in the afore-mentioned sequence of smoking policies the stage of *smoking restrictions* is *omitted*. A *direct* transition from unrestricted smoking to smoking bans would only be possible, if the phase of *medium* level smoking rates were extremely short, as compared to the average

laps of time between two subsequent non-smoking propositions.

- c) The *boundary* between feasible and non-feasible smoking policies is a *monotonously decreasing step-function* of the smoking rate S . As indicated in Fig. 4, smoking policies *below* this step-function are *feasible* and should be observable, when looking at country data. Policies *above* the mentioned line are *not feasible*. If the model proposed in Fig. 3 is empirically correct, there should be no observations in the area of Fig. 4, which is labelled as “Not feasible”. Hence the model is primarily designed for explaining the *non-observables* and much less the observables (cf. Mueller 2012: 81).
- d) If *smoking rates* are *low*, propositions are *never alleviated*: In this situation the policy maker has the power to realize its preferences for strict anti-smoking laws, since according to Fig. 3, their legislation is not inhibited by the political power of the smokers.

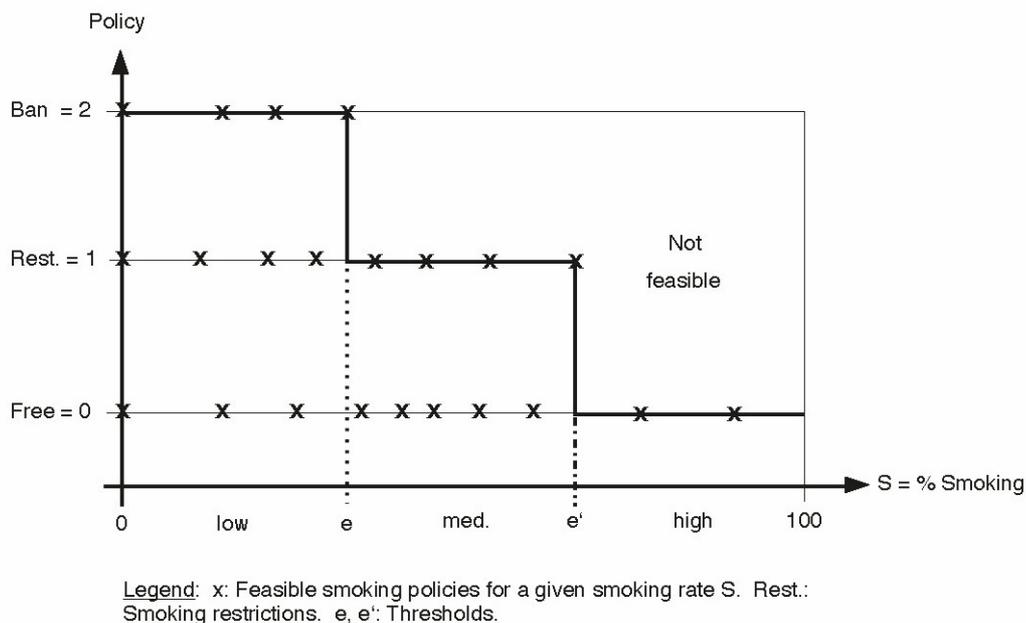


Fig. 4: Feasible and non-feasible smoking-policies: *Theoretical expectations.*

Empirical Validation

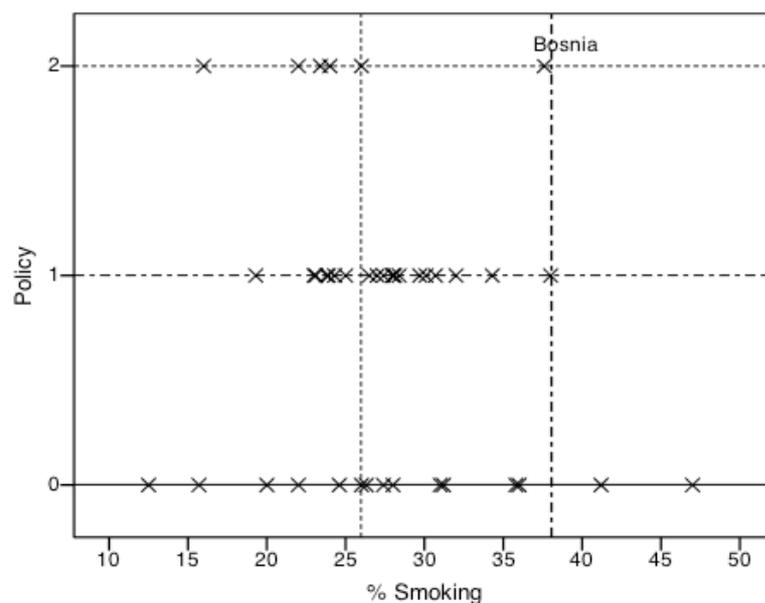
Whenever direct testing of theories is not possible, inferences from associated models are one of the possibilities to corroborate such theories by comparison with empirical data. In principle this also holds for the inferences (a) to (d), which we derived in the previous section from our Petri net model. In practice, however, the appropriate observational data are not always available. This is e.g. the case for the inferences (a) and (b) about

the sequence of anti-smoking policies, which require for testing *time-series* data about the legal development of anti-smoking policies. Even more complicated is the situation with regard to inference (d), which refers to the internal processing of propositions. Since not only the *final* result, but also the *intermediate* stages are here of interest for research, only country based *case studies* of political processes offer sufficient empirical evidence for testing (d). Only inference (c) about the feasibility of

smoking policies as related to the smoking rate S can be tested with the currently available data: the „European Tobacco Control Report 2007“ of WHO (2007) offers systematic country data about the share of smokers as well as the national policies with regard to smoking in restaurants, both for the period around 2005. The policy-data (WHO 2007: 40 ff., tab. 6, col. „Restaurants“) about „free“, unrestricted smoking have been recoded for this study such that they include also „voluntary agreements“, since the latter category obviously bypasses the official law making process. Regarding the share of the smokers S , we refer to the *national* definitions (WHO 2007: pp. 147 ff., annex 4, col. „Total 2002-2005“): they are more heterogeneous than the standardised definitions of WHO (2007: 143 ff.) but also more relevant for national legislators, who always have to consider their own re-election by smokers and non-smokers, when designing a new smoking policy.

In order to test the mentioned inference (c), we plotted in Fig. 5 the %-share S of smokers against the national smoking-policies for restaurants. If the hy-

pothesis (c) were true, we should observe the data-pattern of the previous Fig. 4. This is indeed the case, if we set the thresholds e and e' of the step-function equal to the highest values among those percentages S of Policy=1 and Policy=2, which were in our box-plot analyses *not* yet statistical outliers (Fielding and Gilbert 2000: 127). As theoretically expected, $e'=38\%$ smokers marks the threshold, above which neither smoking-bans nor restrictions are possible (see Fig. 5, dash-dotted lines), probably because of the strength of the tobacco-lobby (Dearlove et al. 2002). Right of the other, dotted threshold indicating $e=26\%$ smokers, restrictions are possible up to $e'=38\%$, but smoking-bans are not feasible. There is, however one exception to this empirical regularity: *Bosnia*, which is a *statistical outlier* far right of the other observations with Policy=2 (see Fig. 5). Its statistically exceptional situation may have to do with the rather late independence of this country in 1997. It compelled Bosnia to make new laws, also with regard to tobacco consumption, which fitted to the health-oriented spirit of that time.



Legend: Policy = 0: „Free“ or „Voluntary agreement“; Policy = 1: „Restrictions“; Policy = 2: „Ban“. **Sources:** Policy: WHO 2007, pp. 40 ff., tab. 6, col. „Restaurants“. % Smoking: WHO 2007, pp. 147 ff., annex 4, col. „Total 2002-2005“. **Sample:** All countries from the WHO-region *Europe* with complete data (N=41).

Fig. 5: Feasible and non-feasible policies for smoking in restaurants: *Empirical observations* around the year 2005.

All in all, there is some evidence, that our model is empirically correct. If it were not, the whole square of Fig. 5 would be filled with randomly distributed data-points. However, it is obvious that the corroboration of inference (c) depends on a few data-points and thus has to be further validated by additional empirical evidence, at best on the basis of time series data. This new information could also be used for corroborating the infer-

ences (a) and (b) about the sequence of anti-smoking policies, which could not be tested in this article.

Some Policy Implications

As mentioned earlier, in policy analysis models are not only used for explanatory purposes, but also for *social engineering*. In this domain, model exploration means

policy *optimisation*, proof of the *feasibility* of a policy, or anticipation of its *unintended negative consequences*. Provided that the remaining part of the empirical validation succeeds, our Petri net model of smoking-bans and -restrictions may also be used for such purposes. In particular, the inferences previously gained by formal reasoning have two major policy implications:

- a) If the process of health reforms for preventing passive smoking starts too *early*, it may be unfeasible. Thus, the policy making has to be coordinated with another Petri process: the secular decline of the political power of the smokers, who at the early stage of this process could block “premature” anti-smoking laws – with detrimental effects on governmental credibility.
- b) If the mentioned health reforms begin too *late*, the policy maker can be criticised for not doing enough in order to protect the general population against passive smoking. Hence, critiques of the governmental health policy could „feed“ the model in Fig. 3 with artificial tokens and demonstrate that smoking-bans or -restrictions would have been feasible even long time before.

In sum: timing and sequence of policies matter and decision support from Petri net models can help to do the right thing in the right moment.

Alternative Models

In view of the importance of the share *S* of the smokers as an exogenous input of the previous Petri net model, one might suppose that non-linear statistical *regression* could be useful not only for estimating some of its parameters but also for replacing the whole model.

The advantage of regression models is indeed their flexibility with regard to the theories, on which they can be based on. Regression lines may e.g. represent an equilibrium derived from a balance of power theory or a maximum resulting from a rational choice approach. The common trait of most of these regression models is however a *deterministic* relation between the dependent and the independent variables, often represented by a linear or a logistic function. Unfortunately it is evident that the spatial distribution of the observations in Fig. 5 cannot be modelled by the mentioned deterministic regression lines. The proposed semi-quantitative Petri net model that distinguishes by a step-function between feasible and non-feasible policies (see Fig. 4) is obviously superior to the regression approach: whereas the *regression model* postulates that at a certain stage of development represented by the share *S* of smokers a certain policy-change *deterministically* occurs, the semi-quantitative *Petri net model* hypothesises that this change can only occur at an *indeterminate* moment after the mentioned stage of development has been reached.

The proposed Petri net has the additional advantage that it allows to formulate and test also hypotheses

about the *sequential transformations* of tokens representing political propositions, which can hardly be modelled with conventional regression equations: one of these hypotheses, which follows from our Petri net model is the *sequence* from unrestricted to restricted and finally banned smoking in public places (see section *Some Inferences from the Model*, inferences (a) and (b)).

SUMMARY AND OUTLOOK

By this article we wanted to explore the usability of the instruments of Petri nets for a new purpose, i.e. *public policy analysis*. Since Petri nets were originally developed for purely technical problems, we had to adapt this toolkit to the needs of *social* modelling. After the introduction of time-dependent white-coloured tokens, inhibitors, triggers, transfer-priorities, and treatment-capacities, we had a modified toolkit for describing both, *qualitative* and *quantitative* aspects of *concurrent* social processes. We successfully used it for an exemplary investigation of smoking-bans and -restrictions. This example is, however, insofar incomplete as it focuses on *policy making* and the explanation of social phenomena. In spite of its intellectual relevance, it neglects the more practice-oriented aspect of the topic: *policy application*, with a strong focus on the analysis of unintended side-effects, process optimisation and the treatment-capacity of the concerned state agencies. All these aspects have not really been discussed in the previous example of *policy making*. Hence, in order to complete the demonstration of the usefulness of Petri nets for public policy analysis, additional examples from *policy application* would be desirable. Since Petri nets have been successfully used for modelling business processes (see e.g. Van der Aalst and Stahl 2011), it is likely that they will also stand this other test.

REFERENCES

- Aldrich, J. and F. Nelson. 1992. *Linear Probability, Logit, and Probit Models*. Sage Publications, Newbury Park.
- Dearlove, J., S. Bialous, and S. Glantz. 2002. „Tobacco Industry Manipulation of the Hospitality Industry to Maintain Smoking in Public Places“. *Tobacco Control* 11, 94-104.
- Dunn, W. 2004. *Public Policy Analysis: An Introduction*. Pearson Prentice Hall, Upper Saddle River.
- Fielding, J. and N. Gilbert. 2000. *Understanding Social Statistics*. Sage, London.
- Goel, R. and M. Nelson. 2008. *Global Efforts to Combat Smoking*. Ashgate, Aldershot.
- Heitmann, F. 2013. „Petri Net Tools Database“. In <http://www.informatik.uni-hamburg.de/TGI/PetriNets/tools/db.html> (accessed at: 1/16/2013).
- Heitsch, S. et al. 2001. „High Level Petri Nets for a Model of Organizational Decision Making“. *Sozionik-aktuell* 1, 17-36. (= <http://www.sozionik-aktuell.de>, accessed at 7/2/2013).
- Jensen, K. 1992. *Coloured Petri Nets*. Springer, Berlin, vol. 1.

- Koehler, M. and H. Roelke. 2001a. „Petrietze als Darstellungstechnik zur Modellierung in der Soziologie (Petri Nets as Tools for Visualising Sociological Models)“. *Sozionik-aktuell* 2, 1-14 (<http://www.sozionik-aktuell.de>, accessed at 7/2/2013).
- Koehler, M., D. Moldt, and H. Roelke. 2001b. „Modelling the Structure and Behaviour of Petri Net Agents“. In *Applications and Theory of Petri Nets 2001*, J.-M. Colom et al. (Eds.). Springer, Berlin, 224-241.
- Mueller, G. 2012. „On the Limits and Possibilities of Causal Explanation with Game Theoretical Models: The Case of Two Party Competition“. *ASK* 21, 69-85.
- Mueller, G. 2013. „„Revolutionen von oben‘ aus der Sicht der Katastrophentheorie: Das Beispiel der Rauch-Verbote in Europa. (‘Revolutions from Above‘ from the Perspective of Catastrophe Theory: Smoking Bans in Europe)“. In *Transnationale Vergesellschaftungen 2013*, H.-G. Soeffner (Ed.). Springer VS, Wiesbaden, CD-ROM, 1-14.
- Newcomer, K. 1994. „Using Statistics Appropriately“. In *Handbook of Practical Program Evaluation*, J. Wholey et al. (Eds.). Jossey-Bass Publishers, San Francisco, chap. 17.
- Parsons, W. 1997. *Public Policy: An Introduction to the Theory and Practice of Policy Analysis*. Edward Elgar, Cheltenham.
- Ragin, Ch. 2007. *The Comparative Method*. University of California Press, Berkeley.
- Reisig W. and G. Rozenberg (Eds.). 1998. *Lectures on Petri Nets I: Basic Models*. Springer, Berlin.
- Rossi, P. and H. Freeman. 1993. *Evaluation: A Systematic Approach*. Sage Publications, Newbury Park.
- Strauss, A. and J. Corbin. 1998. *Basics of Qualitative Research*. Sage Publications, Thousand Oaks.
- Van der Aalst, W. and Ch. Stahl. 2011. *Modeling Business Processes: A Petri Net-Oriented Approach*. MIT Press, Cambridge (Mass.).
- Wang, J. 1998. *Timed Petri Nets: Theory and Application*. Kluwer, Boston.
- WHO. 2007. *The European Tobacco Control Report 2007*. WHO Publications, Copenhagen.

AUTHOR BIOGRAPHY

GEORG P. MUELLER has a Ph.D. from the University of Zurich (Switzerland), where he studied sociology, mathematics, and philosophy. He currently works as senior lecturer (maître d'enseignement et de recherche) at the University of Fribourg (Switzerland), where he teaches research methodology, statistics, and social policy. His research interests include social policy analysis, the construction of social indicators for social monitoring and early warning, as well as the mathematical modelling of social processes.

THE ASSOCIATION BETWEEN GROUP SIZE AND COMMUNICATIONAL COMPLEXITY ACCORDING TO CONCEPTUAL AGREEMENT THEORY

Enrique Canessa
Facultad Ingeniería y Ciencias, CINCO
Universidad Adolfo Ibáñez
Av. P. Hurtado 750, Viña del Mar, Chile
E-mail: ecanessa@uai.cl

Carlos Barra
Independent Researcher
Las Pimpinelas 880, Concon, Chile
E-mail: c.barra@vtr.net

Sergio E. Chaigneau & Ariel Quezada
Escuela Psicología, CINCO
Universidad Adolfo Ibáñez
Av. P. Hurtado 750, Viña del Mar, Chile
E-mail: sergio.chaigneau@uai.cl,
ariel.quezada@uai.cl

KEYWORDS

Conceptual Agreement Theory, evolution of concepts, Agent-based modeling, social complexity hypothesis.

ABSTRACT

We model the evolution of concepts, i.e. how members of a social group associate properties to concepts. Our Agent Based Model (ABM) is based on Conceptual Agreement Theory (CAT), which states that individuals can only infer the conceptual state of others when communicating. Through communication agents develop a conceptual structure which is influenced by three variables: the size of the group, the number of possible properties that may describe each concept and the rate at which agents learn. In general, the results show that these three variables non-linearly interact and that the larger the group and number of available properties, and the slower the learning process, the richer the conceptual structure that emerges from agents' interactions.

INTRODUCTION

Robin Dunbar has argued that there is a direct relation between the complexity of social systems and the complexity of the communicative systems that regulates their interactions (i.e., the social complexity hypothesis; Dunbar 1993; Freeberg et al. 2012). Dunbar and his collaborators define complex social systems as those in which individuals frequently interact with many different individuals, and define complex communicative systems as those that contain a large number of structurally and functionally distinct elements (Freeberg et al. 2012). However, communication is not nearly as well defined. In their writings, communication relates to phenomena as different as signaling, sharing information, and classifying individuals into types. In contrast to these outlooks on communication, we have argued elsewhere that communication, at least in humans, can be advantageously viewed as the process of using concepts to infer shared mental content, i.e., to infer agreement (Canessa and Chaigneau 2013; Chaigneau et al. 2012). Our first goal in the current work is to provide computational evidence for the social complexity hypothesis starting from our own framework about communication, and to test its capacity to generate

novel insights and hypothesis that relate group size and the complexity of the conceptual structure used in communication (i.e., the number of independent structural elements).

When studied in actual social groups, concepts used in language have several interesting properties: (1) Concepts used in language can be sufficiently described by a finite set of properties i among a larger but still finite set of possible properties (Hampton 1979; Rosch and Mervis 1975; Rosch et al. 1976; Smith 1978); (2) These conceptual properties are stable in time (McRae et al. 2005; van Overschelde et al. 2004); (3) Though entities may be categorized in multiple manners (D'Lauro et al. 2008; Murphy and Brownell 1985; Patalano et al. 2006; Rogers and Patterson 2007; Rosch et al. 1976), nouns in language offer only a limited number of alternative conceptualizations for entities; (4) The aforementioned conceptual properties are only probabilistically related to their concepts (Chang, et al. 2011; McRae et al. 2005; Wu and Barsalou 2009); (5) Even if individuals in a social group conceptualize an entity similarly, there are differences between individuals in terms of conceptual content, i.e., there will be intersubjective variability (for discussions about variability, see Barsalou 1987, 1993; Converse 1964).

The first three properties probably reflect the conventionality of concepts encoded in language. If these properties were not true, inferring other people's mental content would become an intractable problem. The fourth and fifth properties, particularly the last one, have created difficulties for researchers and thinkers on the topic. The problem may be summarized in the question of whether we can still say that concepts are shared given that there are no necessary relations between properties and concepts, and given that there is variability in conceptual content from one person to the other (Barsalou 1987, 1993; Converse 1964; Frege 1893/1952; Glock 2009).

Finally note that the above discussion of communication according to CAT implies that meaning is inferred instead of just effortlessly and transparently transmitted among people. Thus, our approach substantially departs from traditional psychological and sociological research on social influence and opinion dynamics such as Social Impact Dynamics (Latane 1996), segregation (Schelling 1978), and group dynamics (McGrath et al. 2000; Vallacher et al. 2002). All that work ignores the nuances

of meaning inference and takes for granted communication among individuals. Although our focus in this paper is Dunbar's social complexity hypothesis (Dunbar 1993; Freeberg et al. 2012), we think that if our results show that analyzing such ideas using CAT offers new insights, then it would be fruitful to apply CAT to reassessing the traditional work on social influence and opinion dynamics, unpacking the transfer of meaning. That could be done by adding to those models, a new layer of meaning inference according to CAT's ideas, which we will present in the next section.

CONCEPTUAL AGREEMENT THEORY

Agreement is an important aspect in the analysis of many social phenomena, such as public opinion, the spreading of rumor, the formation of social and linguistic conventions (Castellano et al. 2009). Conceptual Agreement Theory (CAT; Chaigneau et al. 2012; Canessa and Chaigneau 2013) models an idealized communication event where participants talk about something they cannot ostensibly define (i.e., something they can't point to). Note that many conversational topics would conform to this description, such as beliefs, opinions, situations not currently in perception, and abstractions. We label these as diffuse concepts. CAT assumes that what people do in these situations is to infer agreement, i.e., to infer whether other people's mind-content is similar to their own content or not. The idealized structure of such conversations is the following. Imagine two individuals, *O* and *A*, that are having a conversation about a given entity. Individual *O* believes that the entity is being jointly conceptualized as an instance of concept *C*. Because in principle *A*'s mental content is private, *O* can only infer whether it is true that the entity is also an instance of *C* for *A* or not. To make this inference, *O* observes *A*, and when *A* describes the entity as having a property of type *i*, *O* evaluates if *i* is consistent with *C* in his mind or not. If it is consistent, then *O* infers that *A* is also talking about the given entity conceptualized as *C* (otherwise, disagreement is inferred).

Though it may not be immediately apparent, because of properties four and five discussed above, these inferences of agreement are necessarily probabilistic. Furthermore, if people carried out conversations following the idealized structure of conversation outlined above, they would sometimes be in true agreement (event *a1*) but at other times they would be in illusory agreement (event *a2*). CAT allows the computation of the probabilities for these two kinds of agreements.

These probabilities are conceptually defined as follows. First, the probability of true agreement (symbolized by $p(a1)$) stands for the probability that two agents (*O* and *A*) agree on something given that they have a version of the same concept in their minds. Second, the probability of illusory agreement (symbolized by $p(a2)$) stands for the probability that *O* and *A* agree given that they hold versions of different concepts in their minds. Though CAT can handle cases where properties *i* belong to

concept *C* and *Cn* following any arbitrary probability distribution (by property four), in this introduction we limit ourselves to the case of equiprobable or uniform distributions of conceptual properties. This case should allow the reader to understand CAT's basic ideas (for a more in depth discussion, we refer the reader to Canessa and Chaigneau 2013, and Chaigneau et al. 2012). For equiprobable cases, we have shown (not included here due to space restrictions) that for concept *C*:

$$p(a1) = \frac{s_1}{k_1} \quad (1)$$

and that

$$p(a2) = \frac{s_1 u}{k_1 k_2} = p(a1) \frac{u}{k_2} \quad (2)$$

where,

k_1 = the total number of properties for a concept *C* in a population of individuals.

s_1 = the average number of property types coherent with concept *C* in an individual's mind ($s_1 \leq k_1$).

k_2 = the total number of property types in an alternative conventional conceptualization *Cn*.

s_2 = the average number of property types coherent with concept *Cn* in an individual's mind ($s_2 \leq k_2$).

u = the number of property types that are consistent with *C* and with *Cn* (i.e., the cardinality of the $C \cap Cn$ set).

To help the reader understand the application of expressions (1) and (2) to compute $p(a1)$ and $p(a2)$, let's imagine the two concepts depicted in Figure 1, where concept *C* includes properties 0 to 4, and concept *Cn* includes properties 3 to 6, with properties 3 and 4 belonging to both concepts.

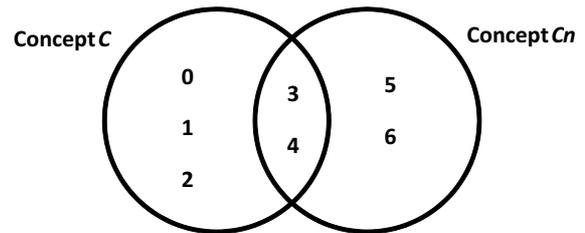


Figure 1: Two Concepts *C* and *Cn* with Their Respective Properties

From Figure 1, we can see that $k_1 = 5$, $k_2 = 4$ and $u = 2$. Assuming that $s_1 = 3$ and $s_2 = 2$, then applying expressions (1) and (2) for concept *C*, $p(a1) = 3/5 = 0.6$ and $p(a2) = (3/5) (2/4) = 3/10 = 0.3$. Similarly for concept *Cn*, the probability $p(a1) = s_2/k_2 = 2/4 = 0.5$ and $p(a2) = (s_2/k_2) (u/k_1) = (2/4) (2/5) = 1/5 = 0.2$. The meaning of these probabilities is easily spelled out by translating them into their conceptual definitions. Thus, $p(a1)$ for *C* means that, in a social group where members know on average s_1 properties of k_1 potential properties for concept *C*, the probability that a given member will find confirmatory evidence for her current

conceptualization C , where this information is being provided by another group member whose current conceptualization is also C , equals 0.6. Similarly, $p(a_2)$ for C means that, in a social group where members know on average s_1 properties of k_1 potential properties for concept C , and where there is an alternative Cn conceptualization with k_2 potential properties that partially overlaps with C by a given u number of properties, the probability that a given member will find confirmatory evidence for her current conceptualization C , where this information is being provided by another group member whose current conceptualization is not C but Cn instead, equals 0.3. An analogous translation can be made for concept Cn .

CONCEPTUAL DESCRIPTION OF THE MODEL

It seems obvious that concepts must be learned. Learning those concepts is part of what makes people members of a given social group. Interestingly, because people are exposed to different experiences, it is likely that they will end up with somewhat different conceptualizations (i.e., there will be intersubjective variability). Furthermore, in contrast to learning a fact of the matter (e.g., that dogs bark), learning about what we call diffuse concepts (as discussed above) arguably requires learning from others whatever is associated with the concept in question. Thus, our Agent-based Model (ABM) always starts a simulation run with agents needing to learn the available concepts. In the agents' environment, there are only two concepts (C and Cn) and a number of potential properties for those concepts (the same properties are potentially associated with both concepts). These two concepts and potential properties represent that concepts are learnable for all agents (i.e., all agents have access to the concepts' potential properties), and provide only a minimal structure so that conceptual structure and agents' agreement may emerge from the agents' interaction history, and not from externally provided information. With regard to those ideas, our ABM may be considered of the bounded confidence type of model (Castellano et al. 2009), focused on the development of diffuse concepts and without imposing a communicational topology among agents. Also, given that CAT states that meaning is inferred between agents, our ABM is similar to the discrimination model (Smith 2001), but much simpler. In Smith (2001), agents use reinforcement learning and keep in mind large discrimination trees, whereas in our ABM only property distributions are kept. Studies have empirically proved that individuals are able to hold in their minds and use frequency distributions (Kane and Woehr 2006; Steiner et al. 1993; Woehr and Miller 1997), which lends face validity to our ABM.

To learn, an Observer (O) agent actively queries an Actor (A) agent by asking it whether property i belongs or not to concept C (or to concept Cn) in A 's mind. To create the query, O randomly chooses either concept C or Cn , and then randomly chooses one property among one of a number of potential properties in its

environment. (For ease of explanation, we will discuss learning for concept C , but exactly the same applies to learning concept Cn .) After receiving the query (e.g., is i a property of C ?), A consults its own conceptual content for the concept in the query and responds. If its concept is empty (as it may occur at the beginning of a run), A chooses a property from the environment, adds it to its concept, and answers the query. If the query is answered positively (i.e., yes, property i belongs to C), O increases the evidential value of property i for the concept in question (i.e., C) in O 's mind. If the query is answered negatively, O decreases the evidential value of property i for concept C . In case the query is answered positively and that property i is already part of the alternative concept (i.e., Cn), O increases i 's evidential value only in half (because it is evidence for two different concepts). The same happens when O decreases the evidential value.

If left unbound, the simple learning mechanism outlined above would probably make all agents eventually learn all the potential properties for the two alternative concepts. Agents would end up with two concepts C and Cn with identical content (i.e., there would not be intersubjective variability). An issue that becomes critical here, then, and that has received little attention in the literature, is whether people stop learning concepts at some point and how do they decide when to stop. If they never stopped, would that lead to everyone having the same conceptual content? In contrast to experimental settings, where people learning concepts are subject to an external standard that will stop learning (e.g., because they achieved asymptotic performance on some learning criterion), in natural settings people must decide for themselves when to stop learning. In our ABM, Observer agents that are learning and that receive positive answers to their queries (as discussed above) decrease the probability of continuing learning, while agents that receive negative answers to their queries, increase the probability of continuing learning. (Increased learning on the face of the unexpected, is a mechanism present in several learning theories, and traceable at least to the Rescorla-Wagner (1972) model of Pavlovian conditioning.) Consequently, the more agents successfully learn, the more probable it becomes that they will use their concepts for communication instead of attempting to learn. When using a concept for communicating, an O agent randomly chooses between C and Cn , and waits for an A agent to produce a property i (note that this is the idealized conversational structure we discussed above). Imagine now that O chooses concept C and that property i produced by A is in fact part of that concept's conceptual content for O . In that case, O would infer agreement, either true or illusory. Furthermore, in case of agreement being found, O increases that property's evidential value for the corresponding concept.

Note that the ABM rules allow agents to stop learning when concepts permit sufficient agreement for communicating, implying that it is not necessary to continue indefinitely learning a concept, and thus

allowing that intersubjective variability occurs. Furthermore, because agents in our ABM learn from other agents in their community, the amount of learning that is necessary depends on variables of the agent group, much as may occur in real social groups (as will be discussed later).

THE ABM'S IMPLEMENTATION AND EXPERIMENT DESCRIPTION

To model the conceptual description, we developed the simplest possible ABM abiding by the *KISS* principle (Keep it simple stupid, Axelrod 1997). In the ABM, akin to Figure 1, there are two concepts C and Cn and P properties that may be part of both concepts. Those properties are represented by numbers from 0 to $P-1$, similarly as depicted in Figure 1. Through the interaction of N agents, agents will form in their minds the conceptual structure for C and Cn , by assigning from the universe of the P properties, some to C and some to Cn . Given that some properties may be assigned by agents to both C and Cn , there might exist an overlap of properties between C and Cn . The assignment of properties to concepts is modeled by P evidential variables $EC_i \geq 0$ and $ECn_i \geq 0$, $i = 0$ to $P-1$, that exist in each agent's mind. A variable EC_i bigger than zero means that the agent has evidence that property i is associated with C in its mind. The same happens with ECn_i for the properties that belong to concept Cn in each agent's mind. The interactions between two agents can be of two types: learning and agreement. The learning interaction models the way agents experience the concepts and learn a concept by assigning properties to concepts and asking for the opinion of other agents regarding that selection. The agreement interaction models how agents communicate among them according to CAT and change their conceptual content depending on whether concepts furnish agreement for communication. The interaction type is probabilistically selected by each agent by means of a learning probability L in the $[0,1]$ interval, which each agent has and will change during the course of a simulation run. At the beginning of a run, given that agents need to learn the concepts, L is set to 1. During a run, that initial value for L will increase or decrease by an amount equal to ΔL , which can be set at a beginning of a run for all agents. In the ABM, agents may represent an Observer (O) or an Actor (A). The following is the description of a simulation step:

1. From the N agents, randomly select without replacement one agent as Observer (O). Then O randomly selects an agent from the rest of the $N - 1$ agents as Actor (A).
2. O probabilistically decides whether to interact with A in learning mode (according to its learning probability (L) that O has in its mind). Thus, O can also decide to interact in agreement mode with probability equal to $1 - L$.
3. Learning interaction:

- a. If in step 2 O decides to interact with A in learning mode, then O randomly selects one of the two concepts (C or Cn), and one property i from the P possible properties and presents that tuple to A .
 - b. If the presented concept in A 's mind does not have associated to it a property (i.e. $EC_i = 0$ or $ECn_i = 0$, $\forall i$, $i = 0$ to $P-1$), then A randomly selects one property i from the P possible properties and assigns it to the chosen concept. That means that A will increase the corresponding EC_i or ECn_i . That increment may be 1 if property i exists in A 's mind only for one of the concepts, or 0.5 if it exists for both concepts. Then, A verifies whether that tuple exist in its mind and communicates that to O . That means that A will check whether the variable EC_i for concept C or ECn_i for concept Cn , for property i , is bigger than zero.
 - c. If the reply from A is affirmative, then O increases EC_i for concept C or ECn_i for concept Cn . That increment may be 1 if property i exists in O 's mind only for one of the concepts, or 0.5 if it exists for both concepts. O also decrements L by an amount equal to ΔL . This means that since the learning activity was successful, the O will increase its probability of acting in agreement mode, i.e. given that O better learned a concept, it will increase the probability of using it for communicating.
 - d. Contrarily, if A 's reply is negative, then O decreases EC_i or ECn_i in the same form already explained. Given that in this case, the learning activity was not successful, the O will increase L by an amount equal to ΔL , which means that it is now more important to O to continue learning the concepts.
4. Agreement interaction:
 - a. If in step 2 O decides to interact with A in agreement mode (with probability equal to $1 - L$), then O randomly selects one of the two concepts (C or Cn) and waits for A to produce a property.
 - b. A randomly selects the concept C or Cn and randomly chooses one property i for the selected concept. If the selected concept in A 's mind does not have associated to it a property (i.e. $EC_i = 0$ or $ECn_i = 0$, $\forall i$, $i = 0$ to $P-1$), then A randomly selects one property i from the P possible properties and assigns it to the chosen concept, which means that A increases the corresponding EC_i or ECn_i by 1 or 0.5 according to the same rule used by A in 3b. Next A presents property i to O .
 - c. O gets A 's property i and verifies whether that property is part of its version of the concept selected by O in 4a. That means that O will see if EC_i or ECn_i is bigger than zero in its mind. If that is the case, O will increase EC_i or ECn_i by 1 or 0.5, according to the same rule used by O

in 3c. In terms of CAT, that means that given that A furnished confirmatory evidence about concept C or C_n , the corresponding property will be useful for communicating in future interactions.

- d. If property i is not contained in the selected concept, then O does nothing. According to CAT, because no agreement was found, O does not strengthen the association of property i with its concept.
5. Repeat steps 1 through 4 until all agents have been O 's.

Though this may not be apparent, the most important difference between learning and agreement interactions is that only in learning mode, and only when A 's reply to O 's query is affirmative, O agents are able to associate new properties with concepts throughout a simulation run. Note, however, this is not true at the very beginning of a run, where A agents are able to relate new properties to concepts in both interaction modes, and without needing a reply from the O agents. This is necessary to kickstart a run, given that A agents initially lack conceptual properties.

The ABM's outputs used in this paper to analyze the model are the following:

1. The total number of properties for concept C in the population of agents (k_1) at the end of a simulation run. That figure is calculated according to CAT. After a run is finished, the ABM counts all the EC_i that are bigger than zero among all the agents. The same for concept C_n (k_2), but counting the EC_n that are bigger than zero.
2. The average number of property types coherent with concept C in agents' minds (s_1). Each agent counts how many EC_i are bigger than zero in its mind and reports that number to the ABM. Then the ABM averages those numbers over all agents. The same is done for concept C_n (s_2).
3. The number of property types that are consistent with C and with C_n (i.e., u , the cardinality of the $C \cap C_n$ set, see Figure 1). The ABM counts the number of EC_i and EC_n that are simultaneously bigger than zero in any agent mind.

We set up the initial learning probability at 1.0 in all experiments, which means that agents begin by learning concepts. This must be so, given that at the beginning of a run, agents do not have any conceptual information in their minds and thus need to learn concepts. Nevertheless, we also used very low initial learning probabilities such as 0.001, and the results that we present here remained the same, which means that results are robust to such parameter. The manipulated input parameters correspond to N (number of agents) and the total number of possible properties (P), which were set at 10, 50 and 90; and the amount by which the learning probability is increased or decreased in O 's minds (ΔL), which was set at 0.05 and 0.2. We performed a $3^2 2^1$ full factorial experiment, comprising

18 different experimental conditions and each condition was replicated 10 times. We chose that design given that we suspected non-linearity in the outputs of the ABM. Values for N and P were chosen to represent small, intermediate and large groups of agents and concepts that could also have a small, intermediate and big number of potential properties, which in this study represent the independent structural elements of the communication system. In general, the more potential properties the system has, the more complex it will be (c.f. Freeberg et al. 2012). Regarding ΔL , we selected a small value for representing a long learning period for agents and a rather high value for modeling a rather short learning period. Finally, the termination condition of a run is defined as the time when the ABM reaches a steady state condition. In that condition, the relevant outputs of the ABM (k_1 , k_2 , s_1 , s_2 , u) remain unchanged. To automatically detect that condition and stop a run, the ABM computes the standard deviation of those figures over a sliding window of 3,000 steps and if all the standard deviations fall below 0.005, it finishes the run.

RESULTS

In the next analyses, we present the results for concept C , given that the results for C_n are analogous. Moreover, remember that we are mainly interested in the $p(a1)$ and $p(a2)$ values and that according to (1) and (2), those probabilities can be calculated using k_1 , k_2 , s_1 , s_2 and u . Thus, most part of the analyses use those probabilities. Figures 2 and 3 present graphs of $p(a1)$ and $p(a2)$ for the 18 experimental conditions. We should note that we performed an ANOVA for the full factorial model and all the components of the model for both $p(a1)$ and $p(a2)$ are statistically significant (p-values ≤ 0.02), except for the $N \times \Delta L$ interaction (p-val = 0.325) and the $N \times P \times \Delta L$ interaction (p-val = 0.225) for $p(a2)$. Although most parts of the model are statistically significant, in this paper we will focus on analyzing main effects and some of the double interactions.

Figure 2 indicates that the smaller the size of the group (small N), the higher $p(a1)$ is. Thus, we can say that small groups will tend to reach a higher true consensus on the meaning of concepts than larger groups. Additionally, the graph suggests that this difference is more noticeable for a bigger ΔL . ΔL regulates how fast the group focuses on agreement interactions and finishes the conceptual learning process, where a smaller ΔL implies a longer learning period.

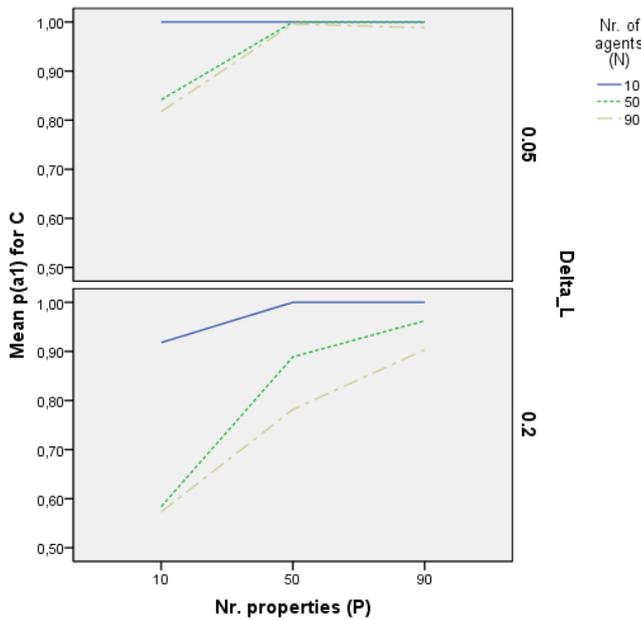


Figure 2: Average $p(a1)$ Values for the 18 Experimental Conditions (avg. over 10 replications for each condition)

Thus, a small Δ_{L} allows agents a longer learning process, asymptotically increasing the number of properties that are assigned to concept C (k_1) from the total number of possible properties (P). The same happens with the size of the average number of property types coherent with concept C in agents' minds (s_1). Therefore, the concept becomes more homogeneous across agents' minds, even if the number of agents increases. Note that since $p(a1) = s_1/k_1$, that explanation is clearly backed up by CAT. The contrary happens with a bigger Δ_{L} . The learning process finishes faster, and thus agents are not able to incorporate many properties in their minds. If the number of agents increases, that situation fosters heterogeneous versions of the concepts in agents' minds. Finally, note that the bigger the universe of properties to describe a concept (P), the higher $p(a1)$. This happens because the greater the number of available properties, the more difficult it is for agents to reach consensus on the set of properties that characterize a concept, and thus the learning period is longer. Therefore, agents are able to describe the concept with more properties, which asymptotically increases k_1 , s_1 and $p(a1)$. Also, a bigger group (larger N) tends to incorporate more properties into concepts (larger k_1), given that more agents interact and each of them can associate different properties to concepts. It is important to mention that, as the ANOVA's significant interactions terms suggest, the effects of the number of agents, Δ_{L} and number of properties on $p(a1)$ interact. For example, more properties slow down the learning process, but that delay is also affected by the Δ_{L} . Also, with a relatively bigger number of agents, Δ_{L} influences the duration of the learning process more significantly than with fewer agents, and correspondingly $p(a1)$, as already explained. The increase in k_1 due to the already explained factors is

analogous to the findings of Lehmann et al. (2011) for cultural traits of a society, where the number of traits increases with population size.

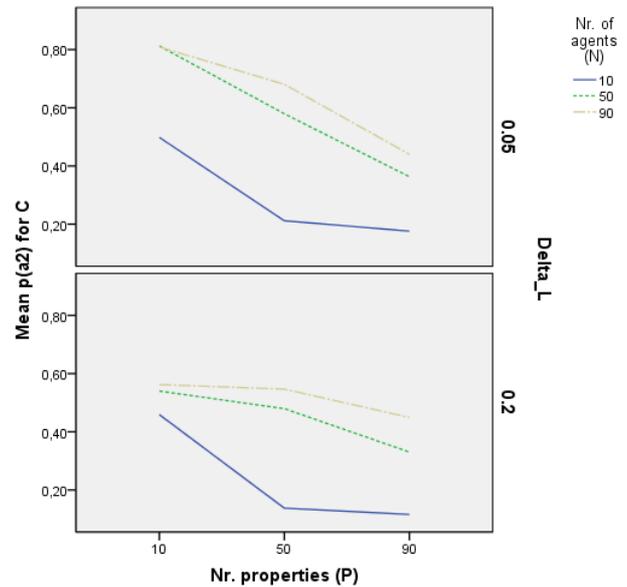


Figure 3: Average $p(a2)$ Values for the 18 Experimental Conditions (avg. over 10 replications for each condition)

In the case of $p(a2)$, Figure 3 shows that the smaller the group, the lower $p(a2)$ is. Given that few agents will have less chance of learning many properties for a concept, then the number of properties that might be simultaneously learned for both concepts will decrease, making u (the number of properties that belong to both concepts) smaller. Given that according to CAT $p(a2) = p(a1) u/k_2$, the smaller u implies a lower $p(a2)$. Figure 4 presents a graph that indeed shows that as the number of agents decreases, u becomes smaller. However that graph also indicates that for a relatively big number of agents ($N = 90$), u does not monotonically decrease as the number of properties increase. Given a small number of properties, necessarily only a few can belong to both concepts, even if many agents interact, i.e. there is a natural limit on the maximum number of properties that can belong to both concepts. In that case note from Figure 4 that $u \approx 10 \approx P$. Now, if the number of properties grows, then a relatively big number of agents can associate more properties with both concepts, and thus u can increase. However, if the number of properties continues increasing, then there are so many available properties for both concepts, that even a large number of agents cannot incorporate all of them in each concept, which makes u decrease. Alternatively, we can say that there is a compromise between the number of available properties and u when there is a large number of agents. Few properties foster a low level of diversity in conceptual learning and thus it is easy for a large number of agents to associate most of those properties with both concepts. On the contrary, if many available properties exist, that promotes a high level of diversity, thus the probability that many of the same properties are

included in both concepts decreases, if the ratio of number of properties to number of agents is too high.

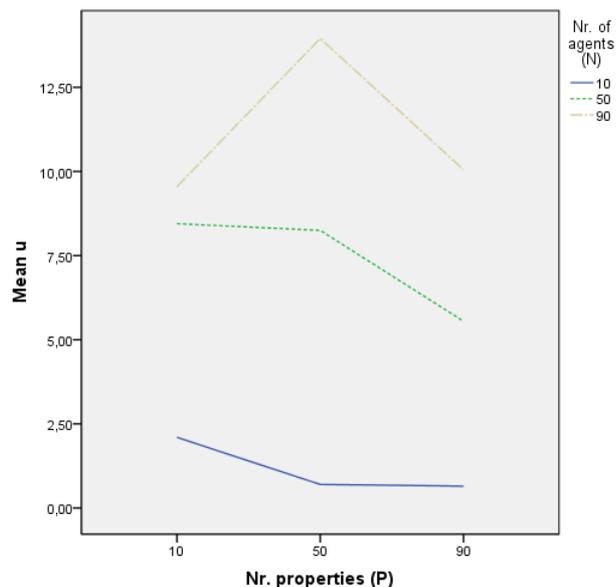


Figure 4: Average u Values for Different P and N (avg. over 10 replications for each condition)

DISCUSSION/CONCLUSIONS

If we jointly analyze the effect of the factors on $p(a1)$ and $p(a2)$, we can draw interesting additional conclusions. Our results are broadly consistent with Dunbar's social complexity hypothesis (Dunbar 1993; Freeberg et al. 2012), but add several interesting nuances. Recall that the social complexity hypothesis states that there is a direct relation between the complexity of social systems and the complexity of the communicative systems that regulates their interactions (i.e., the number of independent structural elements). At a very broad level of analysis, we found that in our simulations, the greater the number of agents, the greater the number of properties that are incorporated into their concepts. This is precisely what the social complexity hypothesis holds.

Now, to the nuances. A first nuance is that our simulations predict that large groups will produce complex communication systems but at the expense of introducing uncertainties in communication (i.e., less true agreement and more illusory agreement relative to small groups). Small groups in our simulations tend to develop quite simple concepts (i.e., with a small k_l and s_l). On the up side, these simple concepts produce rather high $p(a1)$ and low $p(a2)$, meaning that small groups create a conceptual structure that fosters low uncertainties in communication. However, on the down side, this occurs because small groups are fast to close concepts, as soon as those concepts allow agreement. Thus, their concepts reflect more the structure of their common history of interactions than the concept's initial structure (recall that the initial structure stipulates that any property could be associated with any concept). In stark contrast, large groups develop rich concepts (i.e.,

large k_l and s_l) that depend less on the group's history of interactions, but at the expense, as already stated, of relatively lower $p(a1)$ and higher $p(a2)$, while still preserving that $p(a1) > p(a2)$ (which is a requisite for communication).

A second nuance is that our simulations predict that in the real world, the ΔL parameter should have an important influence on the conceptual structure that groups will develop. In the ABM results, ΔL interacts with the number of agents (N) and number of properties (P) for $p(a1)$. Thus, those three variables dictate the size of the set of properties that describe a concept (k_l) and the average number of property types coherent with a concept in agents' minds (s_l). Such a combination of the corresponding real world parameters, will affect the rate at which people learn concepts and when individuals think that they should stop learning. That will in turn influence the richness of the conceptual structure that people may acquire through time. Though the number of properties and the size of the social group are both variables in the social complexity hypothesis, a new idea that our ABM offers is that there is a difference between leaning concepts versus using concepts for communication, and that the complexity of a communication system may depend also on how soon people stop learning and start using concepts (i.e., the ΔL parameter). It is possible that in real social systems many psychological and social variables could influence the ΔL parameter, and thus affect the complexity of concepts used in that group. We believe that the influence of the ΔL parameter in our simulations suggests that for the development of real communication systems, it is necessary that concepts remain open to learning for a prolonged period. Although all the above-discussed conclusions seem intuitively plausible, notice that the results show that the effects of the factors are highly non-linear, with many interactions among the factors. Also, as Figures 3 and 4 suggest, these interactions even change the direction of the curvature of the effects, which is an interesting issue to be investigated. In particular, as already discussed, Figure 4 shows that for large groups, u increases with the number of properties, but then decreases, effect that we think is far from being intuitive.

Now we turn to offer a real world example that could follow a similar dynamics as discussed above. A many times replicated finding is that children from lower Socioeconomic Status (SES) increase their vocabularies at slower rates than children from higher SES, and probably end up with smaller vocabularies overall (e.g., Morrisset et al. 1990). If we take this small vocabulary size to be roughly analogous to a small number of properties in concepts in our simulations, it seems obvious that this phenomenon occurs because, just as in our simulations, the process of learning a communication system will stop when concepts being learned become useful for communication. It's again obvious that this is controlled by characteristics of the environment of the group where learning takes place. If

the environment where a group exists has a small number of potential properties to describe concepts (i.e. a small P), then learning will stop soon and no further properties will become associated with the concept, because there is nothing more to learn. However, our simulations suggest insights further from the obvious, which may even be empirically tested. All other things being equal, SES groups that are larger or better connected should have more complex communications systems (i.e., larger vocabularies, richer concepts). Up to a certain limit, if a social group has more intersubjective variability in conceptual content, then its members should have richer concepts. If children from low SES interact early in life with large and heterogeneous groups, they should develop larger vocabularies and richer concepts. If the ΔL parameter could be somehow manipulated, then people would remain in a learning phase for sufficient time to allow their vocabularies and concepts to be influenced by increasingly large and diverse groups, making their concepts correspondingly richer. We acknowledge that in this example we are equating words (the vocabularies) with conceptual properties, and that this may be questioned. Note, however, that a clear-cut separation between properties and concepts is achieved only at the level of concrete nouns (e.g., it seems perfectly valid to say that *wags its tail* is a perceptual property of the concept *dog*). In contrast, abstract words (which we have labeled here, *diffuse concepts*) have other words and concepts as properties (Recchia and Jones 2012). It is true that the properties of abstract concepts are probably semantic properties rather than plain properties. However, we believe they can be treated the same when modeling communication.

Finally, note that all these conclusions are tentative, pending further validation of the ABM. We have already validated with empirical data CAT regarding the validity of equations (1) and (2), i.e. the calculation of the probability of true and illusory agreement. Thus, given that the ABM is based on CAT and other generally accepted theory, which lend face validity to the model, we believe that the ABM could be plausible enough to be used to conduct “thought experiments” (Axelrod 1997). Hence, one could use this ABM to gain insights into the phenomenon and develop hypotheses that could be then tested through experiments with human subjects. All that is part of our future work with the model, which will also include ABM’s experiments with different types of communication topologies among agents, to further delve into the social complexity hypothesis of communication systems (Dunbar 1993; Freeberg et al. 2012).

ACKNOWLEDGMENT

This work was supported by FONDECYT (Fondo Nacional de Ciencia y Tecnología of the Chilean Government) grant Nr. 1130052 to the first and last author.

REFERENCES

- Axelrod, R. 1997. “Advancing the art of simulation in the social sciences”. In *Simulating Social Phenomena 1997*, R. Conte, R. Hegselmann, and P. Terna (Eds.). Springer, Berlin, 21-40.
- Barsalou, L.W. 1987. “The instability of graded structure: implications for the nature of concepts”. In *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization 1987*, U. Neisser (Ed.). Cambridge University Press, Cambridge, 101–140.
- Barsalou, L.W. 1993. “Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols”. In *Theories of Memory 1993*, A. C. Collins, S. E. Gathercole and M. A. Conway (Eds.). Lawrence Erlbaum Associates, London, 29–101.
- Canessa, E. and S. Chaigneau. 2013. “The dynamics of social agreement according to Conceptual Agreement Theory”. *Quality and Quantity*, 1-21. Article in Press, on-line, DOI 10.1007/s11355-013-9957-7.
- Castellano, C.; S. Fortunato and V. Loreto. 2009. “Statistical physics of social dynamics”. *Reviews of Modern Physics*, 81(2), 591–646.
- Chaigneau, S.E.; E. Canessa and J. Gaete. 2012. “Conceptual agreement theory”. *New Ideas in Psychology*, 30(2), 179–189.
- Chang, K. K.; T. Mitchell and M. A. Just. 2011. “Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fMRI activation”. *NeuroImage*, 56(2), 716–727.
- Converse, P.E. 1964. “The nature of belief systems in mass publics”. In *Ideology and Discontent 1964*, D. E. Apter (Ed.). The Free Press, New York, 206–261.
- D’Lauro, C.; J.W. Tanaka and T. Curran. 2008. “The preferred level of face categorization depends on discriminability”. *Psychonomic Bulletin & Review*, 15, 623–629.
- Dunbar, R. I. M. 1993. “Coevolution of neocortical size, group size and language in humans”. *Behavioral and Brain Sciences*, 16, 681–735.
- Freeberg, T. M.; R. I. M. Dunbar and T. J. Ord. 2012. “Social complexity as a proximate and ultimate factor in communicative complexity”. *Philosophical Transactions of the Royal Society B*, 367, 1785–1801.
- Frege, G. 1893/1952. “On sense and reference”. In *Translations from the philosophical writings of Gottlob Frege 1952*, P. Geach and M. Black (Eds.). Blackwell, Oxford, 56–78.
- Glock, H. J. 2009. “Concepts: where subjectivism goes wrong”. *Philosophy*, 84(1), 5–29.
- Hampton, J. A. 1979. “Polymorphous concepts in semantic memory”. *Journal of Verbal Learning and Verbal Behavior*, 18, 441–461.
- Kane, J. S., & Woehr, D. J. 2006. “Performance Measurement Reconsidered: An Examination of Frequency Estimation as a Basis for Assessment”. In *Performance Measurement: Current Perspectives and Future Challenges, 2006*, W. Bennett, Jr., D. J. Woehr and C. E. Lance (Eds.). Mahwah, Lawrence Erlbaum Associates, NJ, 77–110.
- Latané, B. 1996. “Dynamic Social Impact: The Creation of Culture by Communication”. *Journal of Communication*, 46(4), 13–25.
- Lehmann, L.; K. Aoki and M. W. Feldman. 2011. “On the number of independent cultural traits carried by individuals and populations”. *Philosophical Transactions of the Royal Society B*, 366, 424–435.

- McGrath, J. E.; H. Arrow and J. L. Berdahl 2000. "The Study of Groups: Past, Present, and Future". *Personality and Social Psychology Review*, 4(1), 95–105.
- McRae, K.; G. S. Cree; M. S. Seidenberg and C. McNorgan. 2005. "Semantic feature production norms for a large set of living and nonliving things". *Behavioral Research Methods, Instruments, and Computers*, 37, 547–559.
- Morrisset, D.; K. Barnard; M. Greenberg; C. Booth, and S. Spieker. 1990. "Environmental influences on early language development: The context of social risk". *Development and Psychopathology*, 2, 127–149.
- Murphy, G. L., and H. H. Brownell. 1985. "Category differentiation in object recognition: Typicality constraints on the basic category advantage". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 70–84.
- Patalano, A.L.; S. Chin-Parker, B. H. Ross. 2006. "The importance of being coherent: category coherence, cross classification, and reasoning". *Journal of Memory and Language*, 54(3), 407–424.
- Recchia, G. L. and M. N. Jones. 2012. "The semantic richness of abstract concepts". *Frontiers in Human Neuroscience*, 6:315.
- Rescorla, R. A. and A. R. Wagner. 1972. "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement". In *Classical conditioning II: current research and theory 1972*, A. H. Black and W. F. Prokasy (Eds.). Appleton-Century-Crofts, New York, 64–99.
- Rogers, T.T. and K. Patterson. 2007. "Object categorization: reversals and explanations of the basic-level advantage". *Journal of Experimental Psychology: General*. 136(3), 451–469.
- Rosch, E., and C. B. Mervis. 1975. "Family resemblances: studies in the internal structure of categories". *Cognitive Psychology*, 7(4), 573–605.
- Rosch, E.; C. B. Mervis; W. D. Gray; Johnson, D. M. and P. Boyes-Braem. 1976. "Basic objects in natural categories". *Cognitive Psychology*, 8, 382–439.
- Schelling, T. C. 1978. "Micromotives and Macrobehavior". Norton, New York.
- Smith, E.E. 1978. "Theories of semantic memory". In *Handbook of Learning and Cognitive Processes 1978*, Estes, W.K. (Ed.) vol. 6. Lawrence Erlbaum Associates, Hillsdale, 1–56.
- Smith, A. D. M. 2001. "Establishing Communication Systems without Explicit Meaning Transmission". *Advances in Artificial Life, Lecture Notes in Computer Science* Volume, 2159, 381–390.
- Steiner, D.D.; J. S. Rain and M.M. Smalley. 1993. "Distributional ratings of performance: Further examination of a new rating format". *Journal of Applied Psychology*, 78, 438–442.
- Vallacher, R.; S. J. Read and A. Nowak. 2002. "The Dynamical Perspective in Personality and Social Psychology". *Personality and Social Psychology Review*, 6(2), 264–273.
- Van Overschelde, J. P.; K. A. Rawson and J. Dunlosky. 2004. "Category norms: An updated and expanded version of the Battig and Montague (1969) norms". *Journal of Memory & Language*, 50, 289–335.
- Woehr, D. J., and M. J. Miller 1997. "Distributional Ratings of Performance: More Evidence for a New Rating Format". *Journal of Management*, 23(5), 705–720.
- Wu, L. and L. W. Barsalou. 2009. "Perceptual simulation in conceptual combination: Evidence from property generation". *Acta Psychologica*, 132(2), 173–189.

AUTHOR BIOGRAPHIES

ENRIQUE CANESSA is an associate professor at Universidad Adolfo Ibáñez, Chile. He holds a PhD in MIS/CIS (2002), a Certificate of Graduate Studies in Complex Systems (2001) and an MBA (1991) from the University of Michigan, USA. His research interests include the study of organizations, sociology and cognitive psychology using ABM.

CARLOS BARRA holds a postgraduate degree in Integrated Political Science (2005) from the Chilean Naval Academy, an MBA (1999) from the Institute for Executive Development (IEDE - Chile) and a Master of Science in Computer Engineering (1996) from F. Santa María University, Chile. His research interests include the study of complex systems.

SERGIO E. CHAIGNEAU is full professor at Universidad Adolfo Ibáñez, Chile. He holds a PhD in Cognitive and Developmental Psychology (2002) from Emory University, USA, and a Master of Arts (1995) in Experimental Psychology from the University of Northern Iowa, USA. His research interests include the study of causal categorization and intersubjective agreement.

ARIEL QUEZADA is an associate professor at Universidad Adolfo Ibáñez, Chile. He holds a PhD in Psychology (2007) from Universitat de Barcelona, Spain. His research interests include the study of complex systems and collective behavior, culture and mental models, and cognitive psychology using ABM.

The Dangers of Ethnocentrism

Giangiaco­mo Bravo

Department of Social Studies, Linnaeus University

Email: giangiaco­mo.bravo@lnu.se

ABSTRACT

Humans often alter their behavior depending on the opponent's group membership, with positive (e.g., support of same-group members) or negative (e.g., stereotyping, oppression, genocide) consequences. An influential model developed by Hammond and Axelrod highlighted the emergence of macro-level "ethnocentric cooperation" from the aggregation of micro-level interactions based on arbitrary tags signaling group membership. In this paper, we replicated this model and extended it to allow a wider array of possible agents' behaviors, including the possibility of harming others. This allowed us to check whether and under which conditions xenophobia can emerge beside or in alternative to ethnocentrism.

INTRODUCTION

Ethnocentrism, more properly known as intergroup bias, is the widespread attitude of humans to change their behavior depending on the opponent's group membership. It can have positive implications, such as helping same-group members, but also lead to negative behaviors towards out-group members—ranging from prejudice and stereotyping, to oppression, and genocide—which are often popularized as examples of xenophobia (Hewstone et al. 2002). Group membership is usually built upon symbolic markers—easily observable characteristics such as language, clothing, skin color, etc.—that may dramatically influence people's behavior (Kurzban et al. 2001; Richerson and Boyd 2001). When group boundaries are dependent on geographical proximity, these symbolic aspects of social identity may also determine spatial segregation between ethnic groups (e.g., Musterd 2005).

Some years ago, Hammond and Axelrod (2006b) developed an influential agent-based model (hereafter HA model) that was aimed at cast new light on these issues. They suggested to look at the emergence of macro-level "ethnocentric cooperation" from the aggregation of micro-level interactions based on a standard Prisoner's dilemma game (hereafter PD game). Their definition of ethnocentric cooperation was centered on in-group favoritism based on the interplay between arbitrary "tags", signaling group membership, and local interaction. The research showed that, once combined with "population viscosity", tags were sufficient to determine "high levels of individually costly cooperation with only minimal cognitive requirements and in the absence of other, more complex mechanisms", such as the action of a central authority or reciprocity-based strategies (Hammond and Axelrod 2006b, 932).

More recently, Hartshorn et al. (2013) replicated this model and suggested to reconsider Hammond and Axelrod's optimism over the positive effect of ethnocentrism in favor-

ing cooperation. They argued that the most likely alternative of Hammond and Axelrod's ethnocentric cooperation is not the lack of cooperation but "humanitarianism", i.e., the willingness to help others independently on their tags. Although important, this study had several shortcomings. In their analysis of the mechanisms through which ethnocentrism develops, Hartshorn and colleagues failed to highlight that the conditions leading to ethnocentric cooperation were the same producing universal cooperation when the possibility of identifying group membership was ruled out (see Hartshorn et al. 2013, Tab. 3). In other words, the high cooperation levels in the HA model could be more the result of local agent interaction than a by-product of tags (see also Jansson 2013).

A second crucial issue is that, while Hammond and Axelrod's explored only the positive side of intergroup-bias, the reality is full of examples of its dark side: xenophobia. Although xenophobia is not an inevitable result of ethnocentrism (e.g., Cashdan 2001), the two features often go side by side (Brewer 2001; Hewstone et al. 2002). Therefore, the fact that the same model leading to ethnocentrism could also explain the emergence of xenophobic behaviors, if only the authors would have allowed a wider range of strategies to evolve, requires a further in-depth analysis, which is presented in this paper. First, a further replication of the HA model will be developed to disentangle the conditions leading to the evolution of ethnocentric behaviors from the ones favoring to unconditional altruism (Study 1). Then the assumption limiting the possible actions to positive behaviors will be relaxed and agents will be allowed to actively harm others (Study 2). This will allow to check whether and under which conditions xenophobia can emerge beside or in alternative to ethnocentrism.

STUDY 1

In this study, we replicated the HA model and extended the analysis originally done by the authors to include no-tag and random-location conditions. Note that these possibility were studied by Hammond and Axelrod using a slightly different model in a paper published in the *Theoretical Population Biology* (hereafter TPB) journal (Hammond and Axelrod 2006a), but neither presented nor discussed in their *Journal of Conflict Resolution* (hereafter JCR) article (Hammond and Axelrod 2006b).

In the HA model, agents are located on a space formed by a $N \times N$ toroidal lattice, meaning that the borders wrap around such that each site on the lattice has exactly four neighbors, i.e., the agents located on the four adjacent North, South, East and West locations. Agents can only interact with their neighbors. In the interaction phase, they decide whether to pay a cost (of giving help) $c > 0$ to give a benefit (of receiving help) $b > c$ to their neighbors, taking a

separate decision for each neighbor. Note that this structure of the game is equivalent to a standard PD game. Therefore, we have labeled the two possible actions of giving or not giving help as cooperation (C) and defection (D) respectively. Costs and benefits are then subtracted/added to the basic “potential to reproduce” (PTR) of each agent, i.e., its probability to produce an offspring in a given period, which is set at a fixed initial at the beginning of the period and is then increased by b each time the agent receives help while it is decreased by c each time it helps.

Each agent i holds an arbitrary tag that was established at its birth and is kept fixed throughout its life. The agent’s strategy was determined by the couple (S_i^s, S_i^d) , where $S_i^s, S_i^d \in \{C, D\}$ with C meaning to cooperate (or “help”) the neighbor and D to defect (or “not help”) it. Agents with $S_i^s = C$ will cooperate with neighbors holding the same tag, while agents with $S_i^d = C$ will cooperate with neighbors holding a different tag; $S_i^s, S_i^d = D$ correspond to defection with agents holding the same and a different tag respectively. Agents can hence follow one of these different strategies: the “ethnocentric” (C, D), the “altruistic” (C, C), the “selfish” (D, D) and the “traitor” (D, C) one.

The simulation runs for T periods. At the beginning of each run, the space is empty. Then in each period the following four stages occur.

1. An “immigrant”—i.e., a new agent—is created in an empty cell with random strategy and tag.
2. Agents’ PTRs are reset to their basic values, then all interactions occur.
3. Agents reproduce with a probability equal to their PTR. Newborns are placed in one of the four neighboring sites, which implies that agents can only reproduce if they have less than four neighbors. Newborns inherit their parents’ characters (S_i^s, S_i^d and tag). However, these can randomly change with probability m .
4. Each agent dies with probability d .

As Hammond and Axelrod (2006b) found that their model results were rather robust to parameter changes, we will present here only results following their standard parameter definition, namely $N = 50$, $T = 2000$, basic PTR = 0.12, $c = 0.01$, $b = 0.03$, $m = 0.005$, $d = 0.1$. The model was implemented in NetLogo (Wilensky 1999).

In order to discriminate between the effect of tag-based strategies and the spatial distribution of agents in the HA model, we proceeded in two steps. First, we compared the outcome of the original configuration with the one obtained by a model where we excluded the tag effect, i.e., where all agents hold the same tag and $S_i^s = S_i^d$. This means that only two strategies are possible: the standard selfish (C, C) and altruistic (D, D) ones. Second, we relaxed the local reproduction rule, i.e., instead of being placed in one of the neighboring sites, newborns can be assigned to any empty patch of the simulated space. This means that four conditions are possible: tags - adjacent location (T-A); no tags - adjacent location (N-A); tags - random location (T-R); no tags - random location (N-R). Being based on a 2×2 full factorial design, this analysis allowed to estimate not only the main effects of tags and localization (and to compare their relative strength) but also the effect of the interaction between the two variables.

Results

For all parameter configurations, we performed 50 runs of the model recording the distribution of agent strategies and their moves in each period. The proportion of cooperative moves in the last 100 periods of the T-A condition, when the system settled to its equilibrium, was 0.80 ± 0.0003 , with the ethnocentric strategy followed by 84% (Tab. I). This was even higher than Hammond and Axelrod (2006b) findings (76% of ethnocentric agents), but it should be noted that their sample was limited to 10 repetition, so that this difference could be due merely by chance. On the other hand, this figure is somewhat lower than the 89% one presented in the TPB paper (Hammond and Axelrod 2006a, Tab. 1). Similarly, the N-A condition led to a high proportion of cooperative moves (0.86 ± 0.0003), with a large dominance of altruistic strategies (86%).

The outcome dramatically changed when the adjacent location rule was relaxed. In the T-R condition the proportion of cooperative moves was only 0.03 ± 0.0003 , while it was 0.02 ± 0.0001 in the N-R condition. In both cases, selfish strategies largely proliferated. Note also that, given that the number of offspring was no longer limited by the available space in the neighborhood of existing agents, the average number of agents was higher in both random location conditions (Tab. I). In all cases, the dynamics of the simulation led to an early establishment of one of the strategies with limited changes afterwards (Fig. 1).

Regressing the simulation conditions on the proportion of cooperative moves led to even clearer results. OLS estimates of a model considering both the main and interaction effects of the two variables showed that, over a background of full defection (except for the randomness introduced by the mutation rate), the introduction of tags increased the proportion of cooperative moves by 0.015 ± 0.003 ($t = 7.88$, $p < 0.001$) while introducing the adjacent location rule increased it by more than 50 times, namely by 0.847 ± 0.003 ($t = 283.70$, $p < 0.001$). Note also that the interaction term estimate had a negative sign (-0.084 ± 0.004 , $t = -19.78$, $p < 0.001$), meaning that tags actually *reduced* the positive effect of adjacent location on cooperation.

Discussion of Study 1 results

Our results showed that tag-based strategies do not favor cooperation *per se*. On the contrary, by discriminating between agents holding different tags, they reduce the potential level of cooperation ideally favored by the adjacent location condition. This is clear contrasting the B and D panels in Figure 2. In the B one, with no tags allowed, altruists form large cluster with only small marginalized groups of selfish agents. On the contrary, in the D panel almost all borders between clusters of agents holding different tags are marked by non-cooperative moves, mainly performed by ethnocentric agents, leading to overall higher defection levels. Note also that the introduction of tags in the random location condition did not significantly improve cooperation. This because in well mixed populations the probability of meeting an agent with a different tag is higher than the one of meeting an agent with the same tag.

Note that, as highlighted by Nowak et al. (1994), spatial proximity allows *per se* selfish and altruistic strategies to

	Adjacent location				Random location			
	Tags		No tags		Tags		No tags	
	mean	sd	mean	sd	mean	sd	mean	sd
(C,C) <i>altruist</i>	0.096	0.032	0.857	0.028	0.005	0.002	0.016	0.005
(C,D) <i>ethnocentric</i>	0.837	0.037	0.000	0.000	0.056	0.023	0.000	0.000
(D,C) <i>traitor</i>	0.016	0.008	0.000	0.000	0.017	0.005	0.000	0.000
(D,D) <i>selfish</i>	0.052	0.016	0.143	0.028	0.922	0.025	0.984	0.005
Average S_i^c	0.932	0.018	0.857	0.028	0.061	0.024	0.016	0.005
Average S_i^d	0.111	0.035	0.857	0.028	0.022	0.006	0.016	0.005
Cooperative moves	0.795	0.021	0.864	0.027	0.032	0.008	0.017	0.005
Number of agents	1583.132	44.250	1609.550	46.439	2272.999	0.042	2273.000	0.000

TABLE I: Study 1 results. The table reports statistics for the last 100 periods of the simulations.

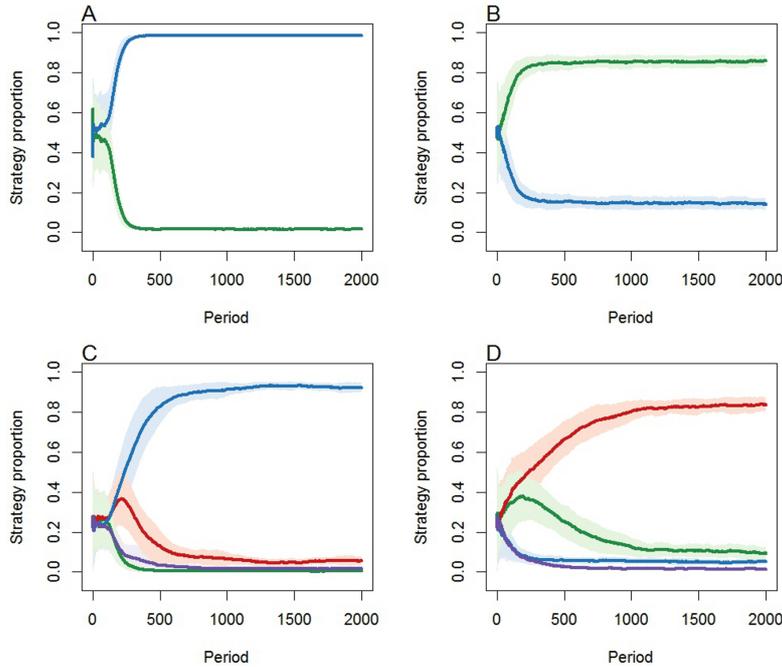


Fig. 1. Dynamics of the HA model. The figure presents the average proportion of each strategy for all model runs. Panels: (A) no tags - random location; (B) no tags - adjacent location; (C) tags - random location; (D) tags - adjacent location. Color legend: green = altruist; red = ethnocentric; blue = selfish; purple = traitor. Colored bands indicate standard deviations.

coexist, with cooperation spreading under a relatively large set of conditions (depending on the b/c ratio and on the probability that a newborn is placed in its parent’s neighborhood). Since these conditions are met in the N-A case, the large share of altruist agents at the end of the simulation are fully consistent with this (Fig. 1B and 2B). More generally, as highlighted also by Axelrod et al. (2004), the evolution of altruism under these conditions can be explained mainly through “inclusive fitness” (Hamilton 1964), which states that natural selection will favor the spread of cooperative behaviors as long as $rb - c > 0$, with r representing the coefficient of relatedness between two agents (obviously one between a parent and its offspring in our case, except for mutation). Given the current parameter configuration, this means that, in absence of tags, altruism will spread if altruists have an average number of cooperative neighbors equal or greater to two: a condition that is respected under the adjacent location, but not under the random location condition.

Hammond and Axelrod (2006a) actually acknowledged this in their TPB paper, although arguing that the joint mechanism of tags and population viscosity “vastly increases the range of environments in which contingent altruism can

evolve”. Moreover, in their JCR paper they do not even mention the possibility that cooperation evolves as an effect of the spatial proximity in absence of tags (Hammond and Axelrod 2006b), highlighting the role of ethnocentrism in maintaining high cooperation levels in structured populations. However, the closest competitor for ethnocentrism in viscous populations is not selfishness but altruism, with the former being the main cause of the decline of the latter (Hartshorn et al. 2013). This means that, ethnocentrism should be seen more as a factor *reducing* cooperation rather than something contributing to its development. To sum up, Hammond and Axelrod (2006b) paper had the clear merit to show the mechanisms through which ethnocentrism is likely to emerge and persist in structured populations. At the same time, their claim that “in-group favoritism can be an undemanding yet powerful mechanism for supporting high levels of individually costly cooperation” should be reconsidered as the same conditions they studied are likely to produce cooperation also in absence of ethnocentric strategies.

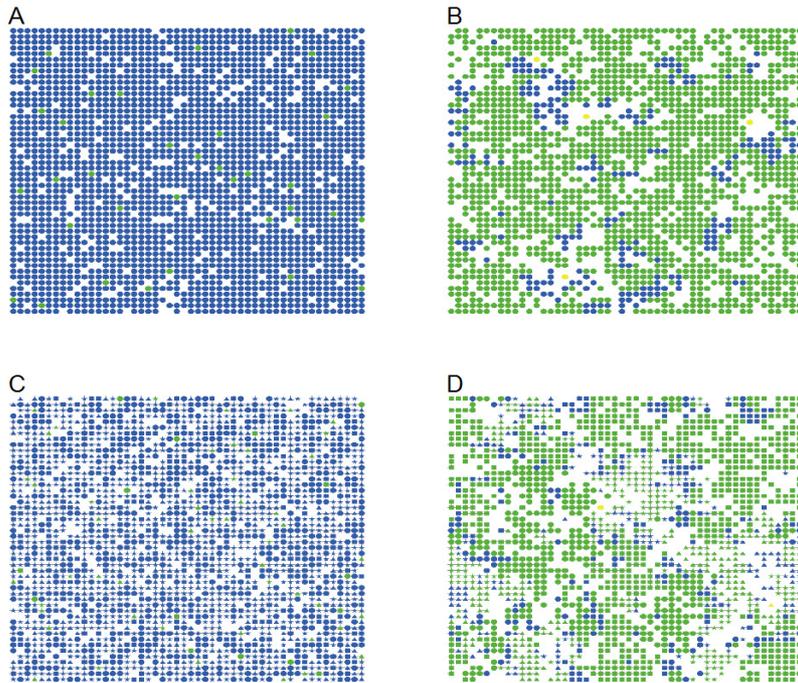


Fig. 2. View of the interaction space at the end of a typical run before reproduction and death under all parameter configurations. Note that only the last move of each agent is shown. Panels: (A) no tags - random location; (B) no tags - adjacent location; (C) tags - random location; (D) tags - adjacent location. Agents' tags are represented by symbols, agents moves by colors: green = C and blue = D . Yellow agents are isolated ones that did not interact in the current period.

STUDY 2

The effect of intergroup bias is not always limited to restraining to cooperate with out-group individuals, but can also take the form of actively harming who is not recognized as a group member (Brewer 2001; Hewstone et al. 2002). This is especially relevant for the evolution of human societies, as warfare among tribes was widespread among hunter gatherers, and ethnic clashes are widespread also today (e.g., Bowles 2012; Keeley 1996; Richerson and Boyd 2001).

A straightforward extension of the HA model is to allow agents not only to benefit, but also to impose a cost on their neighbors. This represents any action harming other people, independently of its specific content (physical attack, limits to personal freedom, etc.). Formally, let us think of a game where, besides choosing as above between helping (C) or not helping (D), players have the possibility to bear a cost c to reduce by b their opponents' PTR, with $b > c > 0$ as before. We called this third option "harming" (H), as it implies an actual reduction the reproduction probability of the opponent. The payoff table of the harming game is presented in Figure 3. Note that the PD game used in Study 1 is a sub-game (given by the four upper-left cells) of the harming one. At the same time, the only Nash equilibrium of the new game is mutual defection, just as in the PD one.

In the harming game model (hereafter HG model), a player's strategy is defined, as before, by the couple (S_i^s, S_i^d) . However, now S_i^s and S_i^d take the values $\{C, D, H\}$ and nine possible strategies exist. Among them, especially important is the (C, H) one, which makes agents both to help those sharing the same tag and to harm agents with different tags. We called this new strategy "xenophobic" (see Hewstone et al. 2002), while (C, D) , (C, C) , (D, D) and (D, C)

are called ethnocentric, altruistic, selfish and traitor respectively, as in the previous study. Except this, everything was equal to Study 1.

Results

We performed 50 runs of the HG model for each of parameter configuration (the same used in the first study), recording in every period the distribution of agent strategies and their moves. As before, the proportion of cooperative moves was considerably higher in the adjacent than in the random location conditions, with the introduction of tags making little difference (Tab. II). However, the distribution of agent strategies was different in the two conditions. While unconditional cooperation dominated in the N-A condition, the modal strategy was the xenophobic one in the T-A condition. Note also that the (D, H) strategy was the only challenging the dominance of altruism in the T-R condition (Tab. II and Fig. 4).

As before, OLS estimates showed that, starting from a situation of universal defection (except for the randomness introduced by mutation), the introduction of tags increased the cooperation proportion only by 0.006 ± 0.0004 ($t = 14.79$, $p < 0.001$) while the introduction of adjacent location increased it by 0.841 ± 0.0004 ($t = 2041.07$, $p < 0.001$). Moreover, the interaction of tags with adjacent location produced a small but significant decrease of cooperation (-0.006 ± 0.0005 , $t = -10.14$, $p < 0.001$).

The proportion of harming moves was relatively low under all conditions even if it significantly increased, up to over 5% of all moves, when both tags and adjacent location were present (II). OLS estimates showed that the proportion of attacks increased by 0.009 ± 0.0002 when tags were

		Player 2		
		Cooperate	Defect	Harm
Player 1	Cooperate	$b - c, b - c$	$-c, b$	$-b - c, b - c$
	Defect	$b, -c$	$0, 0$	$-b, -c$
	Harm	$b - c, -b - c$	$-c, -b$	$-b - c, -b - c$

Fig. 3. The harming game

	Adjacent location				Random location			
	Tags		No tags		Tags		No tags	
	mean	sd	mean	sd	mean	sd	mean	sd
(C,C) <i>altruist</i>	0.089	0.039	0.845	0.032	0.002	0.002	0.011	0.003
(C,D) <i>ethnocentric</i>	0.340	0.132	0.000	0.000	0.022	0.019	0.000	0.000
(C,H) <i>xenophobic</i>	0.500	0.129	0.000	0.000	0.003	0.002	0.000	0.000
(D,C) <i>traitor</i>	0.009	0.007	0.000	0.000	0.057	0.079	0.000	0.000
(D,D) <i>selfish</i>	0.021	0.011	0.137	0.031	0.740	0.282	0.979	0.005
(D,H)	0.027	0.013	0.000	0.000	0.151	0.236	0.000	0.000
(H,C)	0.002	0.002	0.000	0.000	0.002	0.002	0.000	0.000
(H,D)	0.005	0.003	0.000	0.000	0.021	0.015	0.000	0.000
(H,H)	0.006	0.003	0.017	0.005	0.003	0.002	0.010	0.004
Average S_i^s	0.917	0.019	0.828	0.033	0.002	0.017	0.001	0.005
Average S_d^s	-0.432	0.139	0.828	0.033	-0.096	0.186	0.001	0.005
Cooperative moves	0.852	0.026	0.852	0.031	0.017	0.006	0.011	0.003
Harming moves	0.054	0.014	0.016	0.005	0.019	0.005	0.010	0.004
N. of agents	1733.168	32.072	1766.670	29.907	2272.990	0.325	2272.996	0.135

TABLE II: Summary of study 2 results. The table reports statistics for the last 100 periods of the simulations.

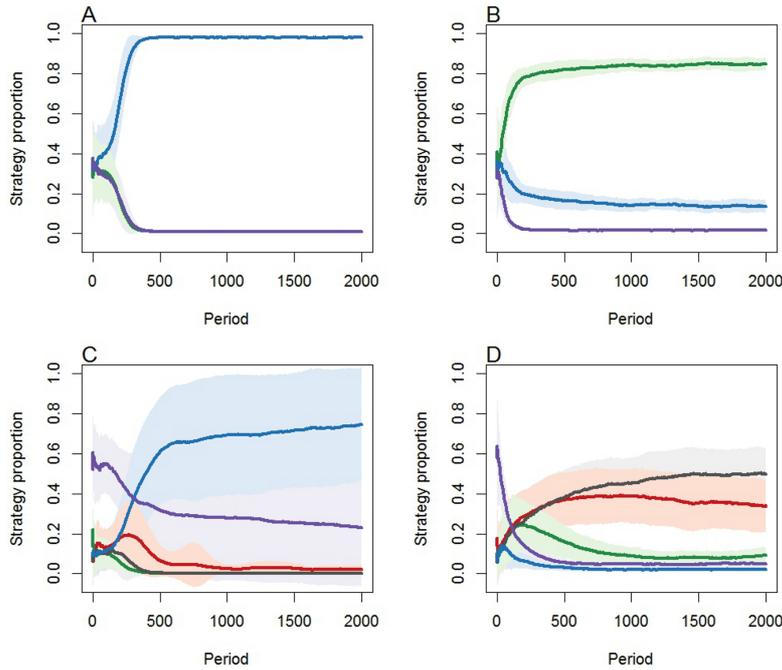


Fig. 4. Dynamics of the harming game. The figure shows the average proportion of each strategy for all model runs. Panels: (A) no tags - random location; (B) no tags - adjacent location; (C) tags - random location; (D) tags - adjacent location. Color legend: green = altruist; red = ethnocentric; gray = xenophobic; blue = selfish; purple = all other strategies. Colored bands indicate standard deviations.

used ($t = 51.67$, $p < 0.001$), by 0.006 ± 0.0002 when adjacent location was introduced ($t = 37.58$, $p = 0.001$), and by 0.029 ± 0.0002 when both factor were at work ($t = 124.73$, $p < 0.001$).

Discussion of Study 2 results

In the HG model, adjacent location was the factor affecting cooperation the most, just as in the first study. Harm-

ing choices were infrequent. Nevertheless, important differences exist between the four conditions. Notably, the fact that harming choices occurred only slightly more often in the T-A than in the other conditions hinders some crucial differences. In the T-A condition, harming occurred infrequently only because it was limited by the presence of well defined borders between the tag-groups (Fig. 5D). On the other hand, the overall harming potential (measured as the proportion of agents with $S_i^s = H$ and, espe-

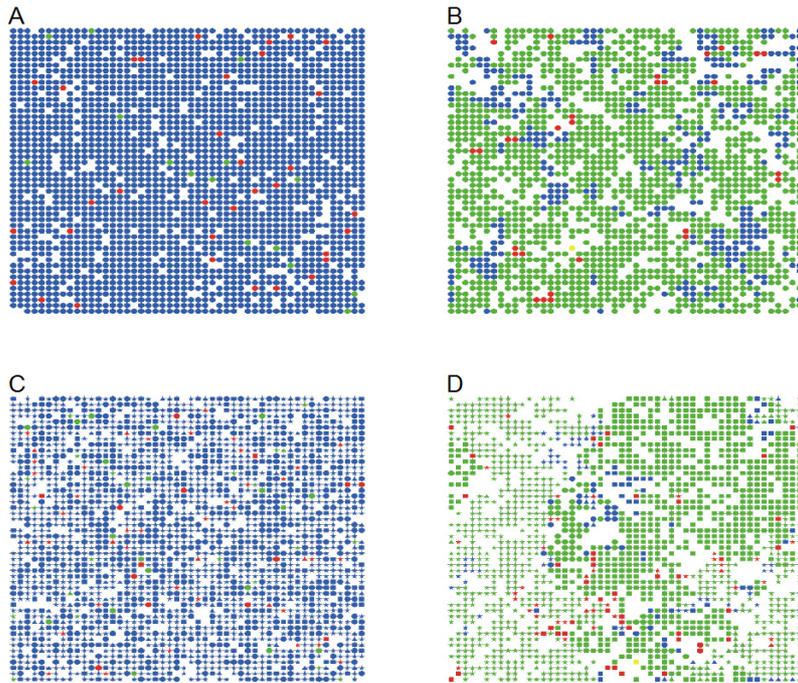


Fig. 5. View of the interaction space at the end of a typical run before reproduction and death under all parameter configurations. Note that only the last move of each agent is presented. Panels: (A) no tags - random location; (B) no tags - adjacent location; (C) tags - random location; (D) tags - adjacent location. Agents' tags are represented by symbols, agents moves by colors: green = C , blue = D , and red = H . Yellow agents are isolated ones that did not interact in the current period.

cially, $S_i^d = H$) was much higher in this condition than in any other one. In other words, while the harming potential of the no-tag conditions is fully exploited, any factor increasing the mixing of agents in the tag ones would have fostered a boost of harming behaviors. This is especially true in the T-A condition, where large cluster of same-tag agents existed that were composed by a majority of agents willing to harm out-group members.

The effects of the harming potential can be revealed by artificially modifying the system once it reached the equilibrium. For instance, changing the agent tags in the T-A condition such that large similar-tag groups were no longer present systematically produced a sudden burst of harming choices, up to 50% of total moves. A similar effect, was also present in the T-R condition, although in the long run this intervention tended to favor the evolution of the selfish strategy at the expenses of the harming ones, (C, H) and (D, H).

Harming strategies evolved mainly because they created extra room for the reproduction of kins (same-tag, same-strategy agents), which means that the spreading of the harming potential depended on the same adjacent location assumption that favors altruism in absence of tags. The effect of this spatial assumption (which may not necessarily hold in the real world) was especially relevant in the T-A condition, where the presence of large same-tag groups allowed borderline agents to afford the extra cost due to their harming actions towards out-group members by means of the cooperation of same-tag agents. In other words, given a sufficiently high segregation, the cooperation of the ethnocentric agents supports xenophobic behaviors, in a situation

of constant between-group conflict which probably mimics human human evolution (Bowles 2012; Keeley 1996).

Finally, it is worth noting that, while in Study 1 ethnocentric cooperation was at least better than universal defection, this was no longer true when considering xenophobic strategies. Consider a simpler situation with only four agents, two of them holding tag A and two tag B. If everyone played selfishly they would obviously get zero in each period. If everyone used the xenophobic strategy, each of them would pay $3c$, once for helping the same-time agent and twice for harming the different-tag agents, while it would receive b because of the help from the same-tag agent and $-2b$ because of the harming from the two different-tag agents. The resulting payoff is $-b - 3c$, which is always smaller than zero. Therefore, the introduction of tags led to a situation that was worst not only than the one where cooperation dominates, but also than the universal defection one.

CONCLUSIONS

Despite the optimistic conclusions of Hammond and Axelrod (2006b), we have shown that the development of ethnocentric strategies might not only limit the potential of cooperation in structured populations, but should be considered as a serious danger due to its "natural" coexistence of this behavior with less peaceful strategies. When the ban of actively harm out-group members is removed, a significant room for the development of xenophobia opens. Moreover, ethnocentric and xenophobic agents easily coexist and mutually reinforce at the expenses of other groups.

The emergence of symbolic markers linked with group distinction should hence be seen more as a potential treat for

the pacific coexistence among individual than as a strategy to increase cooperation in the system. Also spatial segregation looks like a double-edged sword. On the one hand, it can lead to high levels of universal cooperation in absence of tags. On the other, it easily degenerates into parochialism and conflict when group distinctions become explicit (e.g., Bernhard et al. 2006; Choi and Bowles 2007). This is a reason of worry, especially considering that ethnic spatial segregation is widespread in many areas of the the USA and, with some significant national differences, the EU (Musterd 2005).

Fortunately, it must be said that ethnocentrism and xenophobia are not unavoidable. Previous psychological studies showed that common categorizations, e.g., based on race, can be easily overcome by improving the information available to participants (Kurzban et al. 2001). More generally, *decategorization*—i.e., reducing the salience of group distinctions—and *recategorization*—i.e., replacing subordinate (us and them) with superordinate (we) categorizations—can significantly reduce the risks of ethnic clashes (Gaertner and Dovidio 2000; Hewstone et al. 2002). At the same time, a number of other strategies, e.g., institutional development (e.g., Ostrom 2005) and indirect reciprocity systems (e.g., Nowak and Highfield 2011), can be useful to counteract ethnocentrism without condemning societies to ethnic segregation and its risks.

This said, there is also a methodological lesson to learn from this research. While highly abstract models can offer important insights, to take them too literally can be dangerous. Although models, and especially ABMs, can play a relevant role in informing policy making, they need to be specifically designed for the task, and carefully tested against empirical data before advancing any practical suggestion (Boero and Squazzoni 2005; Smajgl and Barreteau 2014). Models are good only as far as their assumptions are tenable and all relevant elements are included. For instance, in the HG model the evolution of harming depended on the adjacent location assumption, which created a competition for space and made the use of strategies creating more room for offspring adaptive. While high competition on a specific resource may or may not be a tenable assumption depending on the case under consideration, to ignore the fact that negative actions against out-groupers often parallel cooperation with fellow group members was clearly a significant limitation of the HA model. As a result, Hammond and Axelrod's optimistic conclusions about the positive effects of ethnocentrism should be seriously put into question.

ACKNOWLEDGMENTS

The author gratefully acknowledges comments offered by Flaminio Squazzoni and by the participants in a seminar at the *Institute For Future Studies* (Stockholm, February 2014) where a previous version of the manuscript was presented.

REFERENCES

Axelrod, R., R. A. Hammond, and A. Grafen (2004). Altruism via kin selection: Strategies that rely on arbitrary tags with which they evolve. *Evolution* 58(8), 1833–1838.
 Bernhard, H., U. Fischbacher, and E. Fehr (2006). Parochial altruism in humans. *Nature* 442, 912–915.
 Boero, R. and F. Squazzoni (2005). Does empirical embeddedness matter?

Methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation* 8(4), 6.
 Bowles, S. (2012). Warriors, levelers, and the role of conflict in human social evolution. *Science* 336(6083), 876–879.
 Brewer, M. B. (2001). Ingroup identification and intergroup conflict: When does ingroup love become outgroup hate? In R. D. Ashmore, L. Jussim, and D. Wilder (Eds.), *Social Identity, Intergroup Conflict, and Conflict Reduction*, pp. 17–41. Oxford: Oxford University Press.
 Cashdan, E. (2001). Ethnocentrism and xenophobia: A cross-cultural study. *Current Anthropology* 42(5), 760–765.
 Choi, J.-K. and S. Bowles (2007). The coevolution of parochial altruism and war. *Science* 318, 636–640.
 Gaertner, S. L. and J. F. Dovidio (2000). *Reducing Intergroup Bias: The Common Ingroup Identity Model*. Philadelphia: Psychology Press.
 Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology* 7, 1–52.
 Hammond, R. A. and R. Axelrod (2006a). Evolution of contingent altruism when cooperation is expensive. *Theoretical Population Biology* 69(3), 333–338.
 Hammond, R. A. and R. Axelrod (2006b). The evolution of ethnocentrism. *Journal of Conflict Resolution* 50(6), 926–936.
 Hartshorn, M., A. Kaznatcheev, and T. Shultz (2013). The evolutionary dominance of ethnocentric cooperation. *Journal of Artificial Societies and Social Simulation* 16(3), 7.
 Hewstone, M., M. Rubin, and H. Willis (2002). Intergroup bias. *Annual Review of Psychology* 53(1), 575–604.
 Jansson, F. (2013). Pitfalls in spatial modelling of ethnocentrism: A simulation analysis of the model of hammond and axelrod. *Journal of Artificial Societies and Social Simulation* 16(3), 2.
 Keeley, L. H. (1996). *War Before Civilization: The Myth of the Peaceful Savage*. Oxford: Oxford University Press.
 Kurzban, R., J. Tooby, and L. Cosmides (2001). Can race be erased? coalitional computation and social categorization. *Proceedings of the National Academy of Sciences* 98(26), 15387–15392.
 Musterd, S. (2005). Social and ethnic segregation in europe: Levels, causes, and effects. *Journal of Urban Affairs* 27(3), 331–348.
 Nowak, M. A., S. Bonhoeffer, and R. M. May (1994). Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Science USA* 91, 4877–4881.
 Nowak, M. A. and R. Highfield (2011). *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. New York: Free Press.
 Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton: Princeton University Press.
 Richerson, P. J. and R. Boyd (2001). The biology of commitment to groups: A tribal instincts hypothesis. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment*, pp. 186–220. New York: Russell Sage Foundation.
 Smajgl, A. and O. Barreteau (Eds.) (2014). *Empirical Agent-Based Modelling: Challenges and Solutions*. New York: Springer.
 Wilensky, U. (1999). Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

DO EDITORS HAVE A SILVER BULLET? AN AGENT-BASED MODEL OF PEER REVIEW

Juan Bautista Cabotà
Departament d'Informàtica
Universitat de València
Avinguda de la Universitat, s/n
46100 Burjassot-València
Email: juan.cabota@uv.es

Francisco Grimaldo
Departament d'Informàtica
Universitat de València
Avinguda de la Universitat, s/n
46100 Burjassot-València
Email: francisco.grimaldo@uv.es

Flaminio Squazzoni
Department of Economics and
Management
University of Brescia, Via San
Faustino 74/B 25122 Brescia
Email: squazzon@eco.unibs.it

KEYWORDS

Peer review; editor; referee-author matching policy; referee behaviour; agent-based modelling.

ABSTRACT

This paper presents an agent-based model of peer review that looks at the effect of different editorial policies of referee selection. We tested four author/referee matching scenarios as follows: random selection of referees, selection of referees with a similar status to submission authors, selection of higher-skilled and lower skilled referees. We tested these scenarios against three types of referee behaviour, i.e., fair, unreliable and strategic and measured their implications for the quality and efficiency of the process. Results show that in case of randomness of referee judgment, any editorial policy is detrimental for peer review. If referees behave strategically, certain matching policies, such as selecting referees of good quality, might counteract possible bias.

INTRODUCTION

Peer review is essential to guarantee the quality of scientific publications (e.g., Squazzoni 2010). Previous studies have tried to measure the effect of peer review on the quality of publications and referees' reports as well as time and costs of different review processes (Jefferson et al. 2002). Other studies have examined the impact of peer review on authors' satisfaction (Weber et al. 2002), the motivation of referees (Squazzoni et al. 2013) and the editors' approach (Neff and Olden 2006; Kravitz et al. 2010; Newton 2010).

A recent large-scale international online survey (Mulligan et al. 2012) called for initiatives to increase the quality of the process. It also indicated the importance of understanding the effect of referee behaviour under different author-referee matching circumstances, such as junior referees reviewing the submissions by senior authors or judgment bias on submissions by young and non prestigious scientists. An analysis on how papers are assigned to referees in a journal (Hamermesh 1994) suggested that except for a very few superstar authors, editors do not usually match authors with referees of similar quality. Another empirical analysis (Callaham and Tercier 2007) showed that academic rank, formal training or status as principal

investigator could not help to predict higher-quality reviews. Some previous work tested a simple selection mechanism based on disagreement control that allows to reduce bias due to lower-skilled or unfair referees (Grimaldo et al. 2011; Grimaldo and Paolucci 2013).

Unfortunately, there is no systematic study of the editor's role in referee selection. This paper tries to fill this gap by simulating the effect of different kinds of author-referee assignments on the quality and efficiency of the process. On the one hand, our model explores different editorial policies to match referees and authors based on their academic status. On the other hand, it also considers different types of referee behaviour, which may be fair, strategic or random and looks at their implications in terms of productivity, efficiency and resource distribution.

THE MODEL

This section follows the ODD protocol for model documentation (Polhill et al. 2008; Grimm et al. 2010).

Purpose

Our model aims to explore the effects of different referee/author matching policies by journal editors. It also explores interaction effects between these policies and different behaviour of referees. It follows previous studies by Squazzoni and Gandelli (2012, 2013) and Cabotà et al. (2013).

By referee behaviour, we mean the way scientists carry out their reviewing task, i.e. their reviewing effort and their possible intentionality. By analysing different editorial matching policies, we aim to explore possible options that editors might consider to improve the quality of the peer review process.

Entities, state variables and scales

The only agent entity in the model is the scientist. Scientists play one of two possible roles in each simulation step: author or referee. The task of an author is to submit an article with the goal of having it accepted for publication, whereas the task of a referee is to evaluate the quality of the submission she/he is matched to. Table 1 summarizes the attributes that parameterize the characteristics of scientists in the model. The editorial policy is modelled through a pair of state variables referring to the review and publication

processes that are represented by the parameters in Table 2.

Table 1: Scientist Attributes

Attr	Name	Brief description
<i>Id</i>	Identifier	Unique agent identifier
<i>R</i>	Resources	Amount of resources accumulated by scientists. They are a proxy of their academic status, position, experience and scientific achievement
<i>Rl</i>	Role	Role played by agents: author/referee
<i>B</i>	Behaviour	The scientist behaviour as referee: unreliable, fair or cheater

Table 2: Editorial Policy Parameters

Attr	Name	Brief description
<i>p</i>	Publication rate	Percentage of submissions that are published.
<i>M</i>	Matching policy	The way editors assign referees to authors: random, similar quality, higher-skilled or lower-skilled.

The model includes different spatial and temporal scales. The spatial scale indicates the number of scientists and the temporal scale the number of simulation steps. One simulation step represents one round in the peer review process, i.e. the submission of papers by authors, the review of all submissions and the selection of publications (see detail below).

Process overview and scheduling

Here, we describe the different processes carried out by the model in each simulation step, in which half of the scientists are randomly selected to play the role of author, while the others act as referees.

Matching authors and referees

Authors and referees are matched on a one to one basis. Therefore, multiple submissions and reviews are not possible and the reviewing effort is equally distributed among the population. We consider four different author/referee matching policies (*M*) as follows: random, peer, higher-skilled and lower-skilled referees. By matching policy we mean the way editors assign referees to authors (for details see the submodels section).

Authors' submissions

The quality of an author's submission ($Q \in [0, 1]$) is a random number from a normal distribution $N(E, \sigma_a)$ with the tails cut off. The mean of this distribution, $E \in [0, 1]$, is the expected quality of the scientist, which is dependent on agent resources as shown in Equation 1.

The parameter $v \in [0, 1]$ indicates the velocity at which the expected quality increases with the increase of the agent's resources. For instance, for a value of $v = 0.1$, each agent needs $R = 10$ to reach a medium-sized expected quality ($E = 0.5$).

$$E = \frac{v^*R}{v^*R+1} \quad (1)$$

The standard deviation of submission quality is calculated as a proportion of the expected quality using Equation 2, where the parameter $av \in [0, 1]$ indicates the level of quality variability when preparing a submission. Indeed, while top scientists could write average or low quality submissions, an average scientist can also write good submissions.

$$\sigma_a = E * av \quad (2)$$

Note that, by calculating σ_a as a proportion of the author's expected quality, the variability in the quality of submissions depends on the scientist's resources (e.g., it will be higher for well established scientists). Indeed, as senior researchers are usually involved in several research lines (that can be at different phases of development) and collaborate with heterogeneous groups (ranging from PhD students to other senior colleagues), it is reasonable to assume that they can produce submissions of different quality.

Referee behaviour

In the role of referees, scientists can be:

- Unreliable, i.e., they do not take reviewing seriously (e.g. due to lack of time or interest) so that their evaluation does not reflect the actual quality of submissions under evaluation.
- Fair, i.e., they consider reviewing seriously and provide an accurate evaluation, which is likely to approximate the actual value of the submission.
- Cheaters, i.e., they consider reviewing seriously to outperform potential competitors by underrating their submissions, even at their own expenses (i.e. resources spent to justify their negative evaluation).

Submission evaluation

To evaluate a submission, referees first estimate the authors' amount of resources. Although a referee can not know the actual amount of resources of an author (R_a), we assume that she/he can have a perception of this, which we called R'_a . R'_a is calculated as a sample value taken from the normal distribution $N(R_a, \sigma_r)$. The mean of this distribution are the author's actual resources, while the standard deviation is a proportion of this mean obtained following Equation 3, where the parameter $rv \in [0, 1]$ indicates the level of variability when perceiving others' resources. This variability mimics the typical knowledge and information asymmetry between authors and referees that characterize peer review.

$$\sigma_r = R_a * rv \quad (3)$$

The quality of the submission as judged by the referee, i.e., $E' \in [0, 1]$, depends on referee behaviour. Unreliable referees can fall into two types of errors: I) recommending submissions of low quality to be published or II) recommending against the publication of submissions which should have been published. We assume that unreliable referees have a fifty percent probability of falling into type I or type II error. If the referee falls into a type I error, E' is calculated as in Equation 4, with $o \in]1, 2]$ as the model's overrating factor. The minimum value of o determines a certain degree of overestimation whereas the maximum would make the referee double estimate the author's resources.

$$E' = \frac{v^*R'_a*o}{v^*R'_a*o+1} \quad (4)$$

When referees fall into a type II error, referees apply an underrating factor to the perceived author's resources as shown in Equation 5. The model parameter $u \in [0,1]$ indicates the percentage of underestimation made by the referee.

$$E' = \frac{v^*R'_a*u}{v^*R'_a*u+1} \quad (5)$$

Fair referees evaluate authors' submission using Equation 6, so that evaluation scores approximate the actual paper quality.

$$E' = \frac{v^*R'_a}{v^*R'_a+1} \quad (6)$$

Finally, referees behaving as cheaters always commit a type II error, as defined for unreliable referees (see Equation 5).

Publication

The publication rate (p) determines the percentage of acceptance so that only submissions getting the highest evaluations are eventually published.

Resource expenses and accumulation

When playing the role of an author, scientists invest all their resources for the submission. If the paper is not published, following the "winner takes all" rule characterising science, authors lose all invested resources. If published, authors accumulate resources according to their publication score, which leads to subsequent submissions of presumably higher quality. The guiding principle is that the more scientists publish, the more resources they have access to, and thus the higher their academic status and position is (e.g., Squazzoni and Gandelli 2012).

We assume that a successful publication increments author resources through a linearly variable multiplication factor in the range $[1, m]$, where m is a parameter of the model. We use higher values for less established authors (i.e. those with less amount of accumulated resources) and we approximate 1 for more established authors. This mimicks reality as publication is crucial in explaining differences in scientists'

performance but is more important for scientists at the initial stages of their academic careers and cannot infinitely increase for top scientists (e.g., Squazzoni and Gandelli 2012).

When acting as referees, scientists allocate resources to reviewing as shown in Equation 7, where $S \in \mathbb{R}$ is the amount of resources that are consumed. The cost of reviewing grows linearly with the quality of author submissions and is proportionally dependent on the referee's resources. If referees are matched with a submission of a quality close to a potential submission of their own, they allocate 50% of their available resources towards reviewing. Accordingly, they spend either less resources when matched with lower quality submissions or more resources when matched with higher quality submissions. Even though top scientists are generally expected to spend less time in reviewing, as they have more experience and are better suited to evaluate sound science than are average scientists, they lose more resources than average scientists because their time is more scarce and valuable (Squazzoni and Gandelli 2012).

$$S = \left(\frac{1}{2}R_r(1 + (Q - E_r))\right)*s \quad (7)$$

Unreliable referees spend less resources than fair and cheating referees as they do not put much effort in reviewing. As indicated by Equation 7, a multiplication factor $s \in [0, 1]$ is applied to reviewing costs. This parameter indicates the percentage of resources saved by unreliable referees, while it is set to 1 in the other cases.

Indexes calculation

In order to measure the quality and efficiency of the peer review process, we define four indexes that are calculated at the end of each simulation step, namely: the evaluation bias, the productivity loss, the reviewing expenses and the Gini index (e.g., Squazzoni and Gandelli 2012).

The evaluation bias (EB) indicates the quality of the peer review process by comparing the optimal situation, in which submissions would have been published according to their quality, to the actual situation in which publication depends on the referee opinion. As shown in Equation 8, we calculate EB as the ratio between the publication errors (PE), i.e. the number of unpublished articles that should have been published and the total number of published articles (PA).

$$EB = \frac{PE}{PA} * 100 \quad (8)$$

The productivity loss (PL) measures the percentage of resources wasted by unpublished authors who deserved to be published. Equation 9 calculates this metric as the percentage of the difference of the quality from the submissions that should have been published (BPQ , from Best Publications Quality) minus the quality from the submissions that were actually published (PQ , from

Published Quality) with respect to the quality from submissions that should have been published (BPQ).

$$PL = \frac{BPQ - PQ}{BPQ} * 100 \quad (9)$$

The reviewing expenses (REx) is the ratio between the total resources spent by referees and the total resources invested by authors. As indicated by Equation 10, they are measured as the sum of resources used by referees in the evaluation process ($RevSp$) divided by the sum of resources invested by authors in their submissions ($AuthSp$).

$$REx = \frac{RevSp}{AuthSp} * 100 \quad (10)$$

Finally, the Gini index measures the inequality in the allocation of resources at the system level; it takes 0 when there is complete equality in the resource distribution and 1 when a single agent has everything. We calculate this index by considering the difference of resources for each pair of agents as in Equation 11, where n is the number of scientists and \bar{R} is the mean of the amount of resources of everyone.

$$Gini = \frac{\sum_{i,j=0}^n |R_i - R_j|}{2 * \bar{R} * n^2} \quad (11)$$

Submodels

Depending on the type of author/referee matching policy and reviewing behaviour, we developed sixteen submodels that include certain typical situations of peer review.

Matching policy scenarios

We distinguish four types of author/referee matching policy (M) scenarios as follows: random, peer, higher-skilled and lower-skilled.

In the “random” matching policy (RMP), authors and referees are randomly matched as if editors would lack knowledge of the scientists’ expertise. This mimics the “luck of the reviewer draw” situation where good quality authors can be matched to lower quality referees and vice-versa. In the other scenarios, we assume that editors have full information on the potential quality of their pool of referees that can be used for referee selection.

In the “peer” matching policy (PMP), authors are matched to referees with similar skills. We arranged authors and referees in two lists that are sorted in descending order by the amount of resources. By following this ordering, authors are paired with their corresponding referees.

In the “higher-skilled” matching policy (HSMP), authors are matched to referees of higher expertise. In this case, for each author, there are two lists of referees: one including referees with equal or greater amount of resources, the other one including referees with lower resources than the author. Then, the author is assigned a random referee from the first list unless there is no one

left (i.e., all of them have been previously matched to other authors). In the latter case, a random referee from the second list is selected.

Finally, in the “lower-skilled” matching policy (LSMP), authors are matched with referees of lower prestige. The logic is the same as in the “higher-skilled” matching policy.

It is worth noting that the last two scenarios mimic situations in which editors could exploit the willingness of high quality scientists to contribute to the reviewing process (i.e. the “higher-skilled” matching policy) or where young scholars (typically PhD students and post-doc researchers) are more frequently involved. It is also worth noting that, in the higher- and lower-skilled scenarios, the success in the application of the matching policy is influenced by the concrete availability of the required referees. For instance, matching an author to a referee with higher resources would not be possible in a situation in which good referees had been already assigned. To check for this, we measured how many times these situations occurred in our simulations. On average, these situations occurred for about 20% of the matchings. On the one hand, this would mimick a realistic constraint of peer review, as editors cannot always find an optimal matching. On the other hand, this constraint has a limited impact on our results.

Reviewing behaviour scenarios

We designed four scenarios for evaluating the effect of unreliable, fair and cheating behaviour of referees.

In the “random behaviour” scenario, there is no room for cheating strategies but only for random behaviour. This is a baseline for studying the effect of the cheating behaviour. During the initialisation of this scenario, we set the scientists behaving unreliably by means of the model parameter $up \in [0, 1]$, which indicates the probability of being unreliable. Then, referee behaviour does not change during the simulation, as if there were no learning or influence from the context.

In the “cheating” scenario, we assume a fixed number of unreliable referees (determined by the up parameter) while the rest of referees behave according to the following criterion: if referees perceive that authors they are matched to have similar or higher resources, they see them as competitors and behave as cheaters; otherwise, they behave fairly. To do so, referees estimate authors’ resources (R'_a) using a normal distribution of the form $N(R_a, \sigma_r)$, where R_a is the actual author’s resources and σ_r follows equation 3.

In the “local competition” scenario, we assume a fixed number of unreliable referees (in accordance with the up parameter), while the rest of referees detect possible competitors only in their own resources neighbourhood. This scenario mimics a situation where referees underrate submissions by authors that have similar resources, while not caring about others. In these cases, their evaluation is fair. To this purpose, we use a Gaussian neighbourhood function of the form $N(R, \sigma_c)$ from which we obtain a sample value Y . The mean of this normal distribution is the referee’s amount of

resources (R_r), while the standard deviation follows Equation 12 using the parameter $cd \in [0, 1]$, which indicates the distance needed to consider an author as a competitor. Then, referees adopt cheating behaviour when the perceived author's resources ($R'_a \in N(R_a, \sigma_r)$) are within the interval $[R_r - (|R_r - Y|), R_r + (|R_r - Y|)]$. Otherwise, referees behave fairly.

$$\sigma_c = R_r * cd \quad (12)$$

In the “glass ceiling” scenario, we assume that there is a fixed number of unreliable referees (again set through the up parameter) while the rest of referees try to outperform both the less and the more productive colleagues. Therefore, given a referee's amount of resources (R_r), the probability of cheating increases when the perceived author's resources ($R'_a \in N(R_a, \sigma_r)$) approach R_r and it is higher when they are greater than R_r . To model the probability of cheating ($P(B=cheating)$), we used the logistic function shown in Equation 13, where d indicates the absolute distance between the author's and the referee's resources (i.e. $d = |R'_a - R_r|$).

$$P(B = cheating) = \frac{1}{(e^{-(\beta_1 * d + \beta_0)} + 1)} \quad (13)$$

Constants β_0 and β_1 determine the shape of the curve as shown in Figure 1 and are calculated as in Equations 14 and 15, respectively. These equations are related to the three model parameters as follows: k_1 indicates the probability of cheating when both referee and author have the same amount of resources (i.e. $d = 0$), and k_2 indicates the probability of cheating when the distance between the author's and the referee's resources is equal to k_3 .

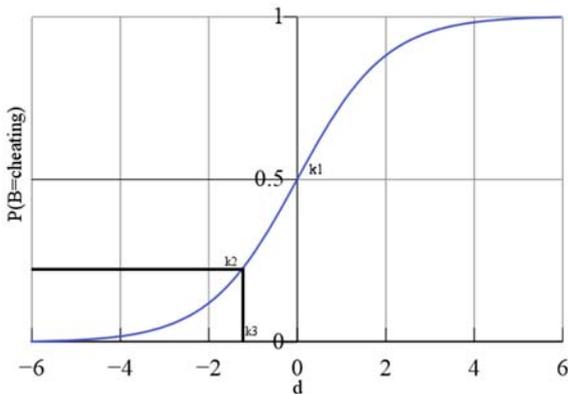


Figure 1: The logistic function used for detecting competitors in the “glass ceiling” scenario.

$$\beta_0 = -\ln\left(\frac{1-k_1}{k_1}\right) \quad (14)$$

$$\beta_1 = \frac{\ln\left(\frac{1-k_2}{k_2}\right) + \beta_0}{k_3} \quad (15)$$

Global parameters of the model

Table 3 shows the model parameters also including the number of scientists (n) and the initial amount of resources for each of them (r_0). It also specifies the range and values that we used to obtain the results shown in the next section.

Table 3: Global parameters of the model

Par.	Description	Range	Values
n	Number of scientists	\mathbb{N}	200
r_0	Initial resources	\mathbb{R}	1
f	Fixed amount of resources gained per agent in each simulation step	\mathbb{R}	1
p	Publication rate	$[0, 1]$	{0.25, 0.5, 0.75}
m	Publication multiplier	\mathbb{R}	1.5
up	Probability of being unreliable	$[0, 1]$	0.5
s	Reviewing expenses factor for unreliable referees	$[0, 1]$	0.5
o	Overrating factor in type I errors	$]1, 2]$	1.9
u	Underrating factor in type II errors	$[0, 1[$	0.1
v	Submission quality velocity	$[0, 1]$	0.1
av	Author variability	$[0, 1]$	0.1
rv	Reviewing variability	$[0, 1]$	0.1
cd	Competitor distance in the local competition scenario	$[0, 1]$	0.5
k_1	Cheating probability when resources are equal in the glass ceiling scenario	$[0, 1]$	{0.75, 0.85, 1}
k_2	Cheating probability when resources differ a value of k_3 in the glass ceiling scenario	$[0, 1]$	{0.4, 0.5, 0.6}
k_3	Resource distance to have a k_2 cheating probability in the glass ceiling scenario	\mathbb{R}	{1, 2, 5, 10}

RESULTS

This section shows the results for all possible combinations of the matching policy and referee behaviour scenarios presented above. For each setting, we averaged the results for 10 runs of 200 simulation steps.

Table 4 shows the results for the most competitive publication rate ($p = 0.25$) which means 25% of submissions eventually published in each simulation step. Other less competitive policies were tested (e.g., p

= 0.50 and $p = 0.75$) that yielded similar results which, for the sake of shortness, we did not report here. In the case of “random” and “local competition” referee behaviour, any editorial matching policy determined more evaluation bias, higher productivity lost and similar or higher reviewing expenses than the random matching. The situation was different in the “cheating” and “glass ceiling” scenarios. In these cases, the “peer” and “higher-skilled” matching policies significantly lowered evaluation bias and the productivity loss compared to the “random” matching policy. In terms of resource distribution, “random behaviour” and “local competition” scenarios generally generated higher values for the Gini index, unless the “peer” matching policy was applied.

Table 4: The effects of editorial matching policies on the peer review process.

	Eval. bias	Prod. loss	Rev. exp.	Gini. index	Cheat. perc.
<i>Random behaviour</i>					
RMP	29.42	15.00	29.42	0.47	NA
PMP	39.55	19.56	34.43	0.37	NA
HSMP	32.99	16.22	30.87	0.43	NA
LSMP	29.51	15.71	29.47	0.46	NA
<i>Cheating</i>					
RMP	70.86	34.72	35.24	0.28	0.27
PMP	51.97	25.69	35.19	0.33	0.25
HSMP	61.95	29.81	34.60	0.30	0.19
LSMP	73.00	36.92	34.86	0.29	0.32
<i>Local competition</i>					
RMP	31.04	15.63	30.13	0.45	0.20
PMP	57.87	28.61	35.70	0.31	0.41
HSMP	36.54	17.74	31.85	0.41	0.22
LSMP	33.47	17.37	30.06	0.44	0.18
<i>Glass ceiling</i>					
RMP	70.35	34.70	34.56	0.29	0.34
PMP	58.02	28.56	35.64	0.32	0.38
HSMP	65.88	32.26	35.23	0.30	0.37
LSMP	68.21	34.47	34.29	0.29	0.36

It is worth noting that the difference of the competitors’ detection mechanism had a considerable effect on the average percentage of cheaters in the population. Generally, this was higher in the “glass ceiling” scenario, although the highest value was reached when the “peer” matching policy was applied in the “local competition” scenario, i.e., when authors were matched with referees of similar quality. Furthermore, it is worth noting that the evaluation bias is not univocally correlated with the number of cheaters in the population. This reflects the importance of the editorial matching policy to influence the peer review process. Figure 2 shows the indexes equilibrium for the “local competition” scenario. In this situation, applying a “peer” matching policy is detrimental as it leads to

higher evaluation bias, productivity loss and reviewing expenses. On the other hand, other editorial matching policies lead to better quality, productivity and efficiency of peer review. Though, this came at a price of having a more unequal distribution of resources, where it is assumed that the best published scientists get the most (e.g., Squazzoni and Gandelli 2013).

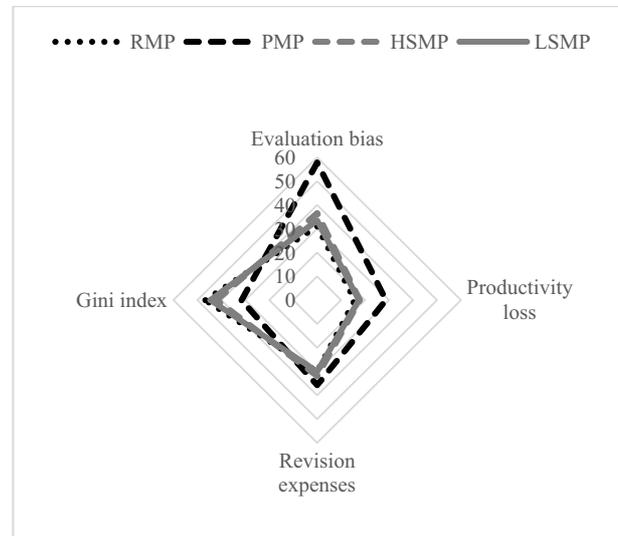


Figure 2: Indexes equilibrium for the different editorial matching policies in a “local competition” behavioural scenario. Note that the Gini index in this figure has been drawn as a percentage.

CONCLUSIONS

This paper aimed to contribute to a growing stream of literature that looks at peer review through agent-based models (e.g., Thurner and Hanel 2011; Allesina 2012; Paolucci and Grimaldo 2014). Given the lack of empirical data on internal processes of peer review, the use of agent-based models can allow us to understand implications of scientist behaviour in idealised situations (Squazzoni and Takacs 2011).

Our results show possible editorial counteractions to reduce the impact of referee misbehaviour, such as matching authors and referees by looking at their reputation. We found that, in case of complete randomness of referee judgment, any editorial matching policy may have even a negative effect. If referees behave strategically, certain matching policies, such as assigning referees of similar or higher quality than submission authors, might counteract referee bias. Besides certain implementation constraints, e.g., the scarce availability of reliable referees, there is also a side-effect of exploiting reliable referees, i.e., generating benefits to published authors who might gain cumulative publication advantages.

Furthermore, we found that peer review outcomes are significantly sensitive to differences in the way scientists identify their competitors. A “man is man’s wolf” competitive scenario increases the chances of

referee bias. Our results showed that certain mechanisms, such as the stratification of scientists in local competing groups and the presence of niches of competition, might reduce the negative effect of cheating and excessive competition. On the other hand, if the competition between scientists is stratified and refers to local groups, the potentially positive effect of editorial matching policies tends to decrease.

REFERENCES

- Allesina, S. 2012. Modeling Peer Review: An Agent-Based Approach". *Ideas in Ecology and Evolution*. 5(2), pp. 27-35. Doi:10.4033/iee.2012.5b.8.f.
- Cabotà, J.B.; Grimaldo, F.; and Squazzoni, F. 2013. "When Competition is Pushed too Hard. An Agent-Based Model of Strategic Behaviour of Referees in Peer Review". *Proceedings 27th European Conference on Modelling and Simulation*.
- Callaham, M. L. and Tercier, J. 2007. "The Relationship of Previous Training and Experience of Journal Peer Reviewers to Subsequent Review Quality". *PLoS Med* 4(1): e40. Doi:10.1371/journal.pmed.0040040.
- Grimaldo, F; Paolucci, M.; Conte, R. 2011. "Agent simulation of peer review: the PR-1 model". *Proceedings 12th International Workshop on Multi-Agent Based Simulation*, LNCS 7124, Pages 1-14.
- Grimaldo, F; Paolucci, M. 2013. "A simulation of disagreement for control of rational cheating in peer review". *Advances in Complex Systems*. Doi: 10.1142/S0219525913500045.
- Grimaldo, F. and Paolucci, M. 2014. "Mechanism change in a simulation of peer review: from junk support to elitism". *Scientometrics*. Doi: 10.1007/s11192-014-1239-1.
- Grimm, V; Berger, U; DeAngelis, D. L.; Polhill, J. G.; Giske, J.; and Railsback, S. F. 2010. "The ODD Protocol: A review and first update". *Ecological Modelling*. Volume 221, Issue 23, 24 November 2010, Pages 2760-2768.
- Hamermesh, D.S. 1994. "Facts and Myths about Refereeing". *Journal of Economic Perspectives*. Volume 8, Number 1, 153-163.
- Jefferson, T.; Wager, E.; and Davidoff, F. 2002. "Measuring the Quality of Editorial Peer Review". *The Journal of the American Medical Association*. Vol. 287, No. 21, 2786-2790.
- Kravitz, R.L.; Franks, P.; Feldman, M.D.; Gerrity, M.; Byrne, C.; and Tierney, W.M. 2010. "Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care?". *PLoS ONE* 5(4): e10072. Doi:10.1371/journal.pone.0010072.
- Mulligan, A.; Hall, L.; and Raphael, E. 2013. "Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers". *Journal of the American Society for Information Science and Technology* 64(1):132-161.
- Neff, B.D. and Olden, J.D. "Is Peer Review a Game of Chance?" *BioScience*, 56(4):333-340, April 2006.
- Newton, D.P. 2010. "Quality and Peer Review of Research: An Adjudicating Role for Editors". *Accountability in Research: Policies and Quality Assurance*. 17:3, 130-145.
- Polhill, J. G; Parker, D; Brown, D.; and Grimm, V. 2008. "Using the ODD Protocol for Describing Three Agent-Based Social Simulation Models of Land-Use Change". *Journal of Artificial Societies and Social Simulation*. Vol. 11, no. 2 3.
- Squazzoni, F. 2010. "Peering into Peer Review". *Sociologica*, 3. DOI:10.2383/33640.
- Squazzoni, F., Bravo, G., Takács, K. 2013. "Does Incentive Provision Increase the Quality of Peer Review? An Experimental Study". *Research Policy*, 42(1), pp. 287-294.
- Squazzoni, F. and Gandelli, C. 2012. "Saint Matthews Strikes Again. An Agent-Based Model of Peer Review and the Scientific Community Structures". *Journal of Informetrics*, 6: 265-275.
- Squazzoni, F. and Gandelli, C. 2013. "Opening the Black Box of Peer Review. An Agent-Based Model of Scientist Behaviour". *Journal of Artificial Societies and Social Simulation*, 16(2) 3: <http://jasss.soc.surrey.ac.uk/16/2/3.html>
- Squazzoni, F. and Takacs K. 2011. "Social Simulation That 'Peers into Peer Review'". *Journal of Artificial Societies and Social Simulation*, 14(4) 3: <http://jasss.soc.surrey.ac.uk/14/4/3.html>.
- Turner, S. and Hanel, R. 2011. "Peer review in a World with Rational Scientists: Toward Selection of the Average". *The European Physical Journal B*, 84, pp. 707-711. Doi: 10.1140/epjb/e2011-20545-7.
- Weber, E.J.; Katz, P. P.; Waeckerle, J.F.; and Callaham, M.L. 2002. "Author Perception of Peer Review". *The Journal of the American Medical Association*. Vol. 287, No. 21, 2790-2793.

AUTHOR BIOGRAPHIES

JUAN BAUTISTA CABOTÀ is PhD. Student at the University of Valencia. His research interests include agent-based modelling and simulation and intelligent decision-making support systems. His email address is: juan.cabota@uv.es.

FRANCISCO GRIMALDO is associate professor at the University of València. His research is focused on agent-based modelling and simulation. He is member of the HIPEAC network of excellence and the IEEE Systems, Man & Cybernetics Society. His email is francisco.grimaldo@uv.es and his webpage can be found at: <http://www.uv.es/grimo/>.

FLAMINIO SQUAZZONI leads the GECS-Research Group on Experimental and Computational Sociology. He is President of ESSA The European Social Simulation Association. His e-mail is: flaminio.squazzoni@unibs.it and his web-page can be found at: www.eco.unibs.it/gecs/Squazzoni.html.

Learning not to Trade: On Scarcity, Emergence and Failure of Markets

Özge Dilaver

Centre for Research in Social Simulation
University of Surrey
24 AC 03 Guildford, Surrey GU2 7XH UK
Email: o.dilaverkalkan@surrey.ac.uk

KEYWORDS

missing markets; scarcity; theory of value; transaction cost; agent-based modelling; zero-intelligence agents

ABSTRACT

This paper addresses the substitution of virgin resources used in industrial processes with by-products that would otherwise be regarded as waste and questions why some by-product markets fail to emerge. It highlights the long-forgotten distinction between the use-value and exchange-value and points out that markets can only work for assigning the latter. It argues that in some industrial settings, waste resources are not scarce and their supply schedule in the classical market plot are at the right of and far from the demand schedule. Hence, even though the waste resource has some positive use-value, a positive price level cannot be established at the intersection of demand and supply schedules. The paper employs agent-based simulations with zero-intelligence traders to illustrate that when agents are unable to learn, transactions do occur in the abovementioned context but when agents have simple adaptive capabilities, the market fails to emerge. Thereby, the paper employs the zero-intelligence traders in a different way than usual; whereas the literature on zero-intelligence traders aim to show markets can emerge even when agents lack rationality, this paper illustrates that markets can fail to emerge when agents have some ability to learn.

INTRODUCTION

Market failure arguments are frequently raised in relation to environmental issues. It is, for example, well known that public goods, such as ecosystem services, can be overused in the absence of institutional arrangements that limit their use or internalise their cost. There is, however, another type of market failure that has important implications for the environment and is not adequately addressed in the existing literature. The non-emergence of some by-product markets is a neglected issue that can lead to 'wasted waste' (Kronenberg and Winkler, 2009); an underuse of waste and corresponding overuse of virgin resources.

The importance of the efficient use of natural resources for current and future generations cannot be overstated. Yet, our production processes usually focus on the value of a small, selected part of a resource and ignore the rest. For example, most agricultural products use only a small part of the crop and even the secondary residues, which are produced while processing the crops and are often available in large quantities at the processing sites, are not fully utilised (Bhattacharya et al, 2005; Johnson and Linke-Hepp, 2007; Gressela and Zilbersteinb, 2007). Secondary forest residues, such as wood bark, is also among the commonly wasted (Anon et al, 1994; Dasappa, 2011; US Department of Energy, 2011). Similarly, some of the waste material produced by the metal processing industries, such as slag and foundry sand, and the coal combustion products, such as ash, can be used in construction industry as substitutes for virgin aggregates. Although these potential uses are well known, most of these materials are stockpiled or sent to landfill (Chertow and Lombardi, 2005; Carpenter and Gardner, 2009). Residual industrial heat is yet another common example of unused potential of waste. While the discharge of residual heat is a clear threat to the environment, wasting important proportions of the energy input via residual heat is common practice in many industrial processes. So much so that cooling towers and smokestacks have become the characteristic elements of industrial landscapes. (Nguyen et al, 2010).

The issue of virgin resource substitution is most frequently addressed in the industrial ecology (IE) literature. This paper, develops a synthesis between the findings of the IE literature and the economic theory on missing markets within a simple theoretical framework. The remainder of the paper is organised as follows. The next section very briefly discusses key issues around missing markets and scarcity to introduce the theoretical background of the paper. The third section presents the lack of scarcity hypothesis in a simple microeconomic framework. The fourth section elaborates on the policy and business implications of the hypothesis focusing on how transactions can be initiated in a lack of scarcity context. The final section presents the conclusions. It may be useful to note that the distinction between waste and by-product emerges through transactions. In order to avoid confusion the term *waste resource* will be used to refer to both waste and by-

product in the rest of this paper.

MARKETS, SCARCITY AND THEIR ABSENCE

Markets are systems of institutions that facilitate and shape exchanges of goods and services. They are complex entities that evolve through path-dependent social processes. A market can be thought to have emerged when agents have an, albeit implicit, understanding of the good and how it will be traded. This understanding is established together with routines and agreements on a broad range of issues such as the essential properties of the good, how its quality can be assessed, the ways in which the buyers and sellers find each other and exchange information, and the type of transactions that are likely to take place. The process by which market institutions evolve may involve elements of learning that reveals what works better in a given context and co-ordination on one of the alternative ways all of which might work. Market institutions typically reflect the nature of the good as well as the broader social context in which the exchanges of the good occur. Shared assumptions about ownership and nature in particular have central roles in the process.

Market failure is the general name for inadequacies of markets in providing economically efficient or socially desirable allocations of scarce resources. It is a negative concept that only makes sense in comparison to a Weberian ideal type of markets that works perfectly. Market failure explains why real life markets do not look or work like textbook descriptions of their ideal types, whether these discrepancies are taken as limitations of the theories themselves, or as lists of what is wrong with the real life markets. A missing market is one kind and, arguably, the extreme case (de Janvry et al, 1991) of market failure. It refers to the non-emergence of the system of institutions that were needed for shaping and facilitating exchanges. Being a type of market failure, the missing market is also a negative concept, which is only identifiable in comparison to a supposedly better alternative. In this case, the benchmark is the potential gains from exchanges at the individual or society level.

Market failure has a special importance in welfare economics because its existence is often taken as an argument for state interference or nonmarket allocation (see Bator, 1958; Arrow, 1970; Barr, 1992; Medema, 2007). Although market failure has been associated with a long list of conditions such as externalities, indivisibilities and public goods, explaining what is at the core of market failure in general or what exactly causes non-emergence of markets appear to be difficult tasks. What emerges from the analyses of Bator (1958), Demsetz (1964), Arrow (1970) and Randall (1983) is that, often, market failure occurs when simple individual ownership of a good does not suffice for realising exchanges that accurately reflect the level of scarcity of that good. Bator (1958) calls this condition a *divorce* between ownership and scarcity.

Hence, scarcity has a central role in both how market failure occurs and why it is important. Within the scarcity-based definition of economics (Robbins, 1945; Samuelson, 1948),

economic goods are, by definition, scarce. *Free goods*, on the contrary, are so plentiful that a particular use of the good does not require diversion from another use. It is useful to make two notes here, as they will be relevant in the following sections. Firstly, from the viewpoint of scarcity-based definition of economics, the cases in which a good is not scarce are trivial and uninteresting to economics. In the following sections, this paper will argue otherwise. Secondly, in this economic conceptualisation of nature and ownership, goods are either zero-priced and freely available in nature for economic agents to exploit, or they are scarce and exchanged at a positive price. A negative price does not have a meaning from this perspective. As it will be explained in the following subsection, however, the common perspective is changing with new conceptualisations of nature.

THE LACK OF SCARCITY MODEL

A. Theoretical Overview:

This section explains the missing markets for waste resources within a simple theoretical framework and an agent-based simulation model. It is not unknown but often neglected that since demand and supply schedules are assumed to be independent from each other, they do not have to intersect at a positive price level. This issue is different from the idea of temporary disequilibria where there is a positive and market-clearing equilibrium price but the transactions occur at a different price, often following an exogenous shock until the necessary adjustments in price, wage and output levels take place.

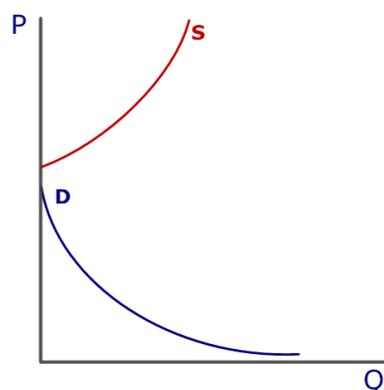


Figure 1:
Arrow's Futures Market

The cases where demand and supply schedule do not intersect are likely to be long-term phenomena that reflect the structural properties of demand and supply and are not expected to disappear through price and output adjustments of individual agents. Arrow's (1970; see also Heller, 1997) explanation for missing markets for some future transactions depicts one such case. Arrow suggests that in the case of futures, it is possible that the highest price any agent is willing to pay for future transactions is still lower than the

lowest price any agent would sell (see Figure 1). It is also recognised that in labour and credit markets (Stiglitz and Weiss, 1981), the supply schedules may be not convex but backward bending (see Figure 2), and do not intersect demand schedules positioned at far right.

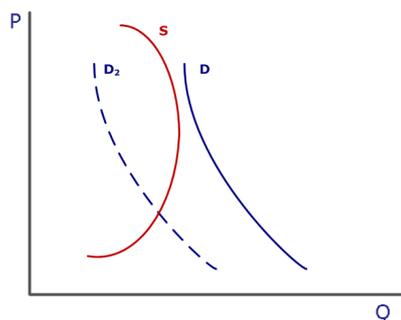


Figure 2:
Stiglitz and Weiss model

B. Assumptions and Justifications:

This paper suggests an explanation that is similar to these seminal contributions for the context of missing markets for waste resources. Here, it is assumed that in many industrial settings, waste resources are not scarce compared to the level of demand for these resources at price levels of zero and above (Assumption 1). Hence, the supply schedule lies at the right of and far away from the demand schedule.

In addition, since the market institutions that facilitate the exchanges have not yet developed in this industrial setting, a lot of innovative, creative and entrepreneurial preparation is needed for establishing expected properties of the good, the unit of exchange and the ways in which the exchanges will take place. It is envisaged that most of these transaction and search costs correspond to technical and organisational preparation, exploring the potential market and contacting potential customers. For this reason, it is assumed that a significant part of transaction and search costs will incur before securing transactions (Assumption 2a).

Finally, it is assumed that the transaction costs are expected to be high, heterogenous and difficult to calculate a priori (Assumption 2b, see Johnstone and de Tilly, 2006, for a review). As producing and exchanging the waste resource is not the main priority or objective of firms, there are likely to be unintended idiosyncrasies between firms in terms of how well and easily they can make those preparations. For example, search and matching costs of different firms may vary due to existing business and social links. The transaction costs would also vary between active and passive suppliers (buyers). If there are spillovers of information, passive suppliers (buyers) may free ride and exploit learning and innovations of others.

These are the main assumptions of the model and they are listed below for clarity.

Assumption 1: Waste resources are not scarce compared to the level of demand.

Assumption 2a: Some of the transaction and search costs occur before the actual exchange takes place.

Assumption 2b: Transaction and search costs are high, heterogenous among firms and difficult to calculate a priori.

Under these revised assumptions, the missing market with lack of scarcity (MMLOS) looks like in Figure 3. The demand and supply schedules do not intersect at a positive price and there is excess supply at each positive price level. Pareto-improving exchanges can occur at infinitely many price levels. Even though the waste resource has some positive use-value, a positive market price cannot be established due to lack of scarcity.

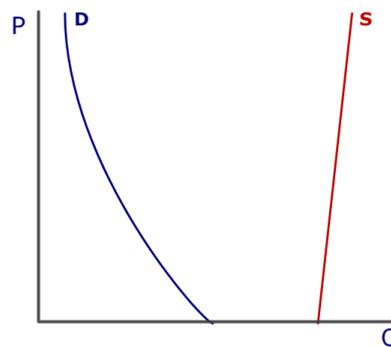


Figure 3:
MMLOS

C. Simulation Model:

I will now explore the properties of the model under different assumptions about agents' rationality and learning. This part of the paper builds on the existing literature on zero-intelligence (ZI) traders following the influential paper by Gode and Sunder (1993). Gode and Sunder conducted experiments comparing human agents with ZI artificial agents in double-auction markets with a central auctioneer who keeps a record of all offers and matches the buyers to sellers. The authors found that the ZI traders can lead to high levels of market efficiency even through random decision making. They concluded that rationality is not a necessary assumption for a market to achieve efficiency; it is rather the market institutions, that is represented by the central auctioneer in their model, produces efficiency.

Thus, the main *problematique* of the ZI traders literature is identifying those capabilities of agents that are necessary conditions of market efficiency given established market institutions. I am, however, referring to a context where such institutions hasn't evolved yet. Thus, the agent-based MMLOS model replaces the central auctioneer with a matching process between agents.

More specifically, the model creates agents in two classes: buyers and sellers. In the experiments reported in this paper, the number of buyers were set as 200 and the number of sellers were set as 30. These agents are assigned demand or

supply schedules that are instances of the market demand and supply schedules respectively. That is, each buyer or seller has linear demand or supply schedules with different slopes and xy intercepts. Aggregation of output levels that agents are willing to buy or sell at each price level given their demand or supply schedules yields the market demand and supply schedules that are in line with an MMLOS context similar to the one shown in Figure 3. This is the basic setup of the simple simulation model used in this study.

The model is then used to run experiments on interactions between traders. At each period, they prepare their offers to buy or sell. Preparation of an offer involves deciding on an output level that they wish to trade and a corresponding price level. Traders decide on the level of output they want to trade, which is a random number between zero and the maximum output they could trade. For buyers, the maximum level of output they can trade is the x intercept of demand curve, i.e. the quantity demanded if the price was zero. For sellers, maximum trade level is the output level they would be willing to sell at a very high price, which is assumed to be slightly above the price of the waste resource's substitute (the virgin resource).

After deciding on the level of output they are willing to trade, the traders establish the price they are ready to accept. This would be the maximum price they are willing to pay if they are buyers and minimum price they are willing to take if they are sellers. Traders establish these prices based on their own demand or supply schedules.

They then announce their offers with output and price level to the market. The matching between buyers and sellers occur randomly under basic simulation settings. If traders can find a trade partner, who is willing to accept their offer, they trade. If the output they traded is lower than their intended level, they can continue making transactions, until this level is reached or there are no available and acceptable offers left in the market.

D. Results of Simulation Experiments:

In the ZI trader literature, constrained traders (ZI-C) know their own demand or supply structure and do not make a transaction that would yield negative profit or utility. The unconstrained traders (ZI-U), on the other hand, would even sell their goods at a loss or buy goods that are overpriced. Studies in this field commonly experiment with different learning capabilities of agents including that of recognising best available price at a point in simulated time.

This study made similar comparisons. In the first set of experiments, the model generates experiments with ZI-U and ZI-C traders, and with and without the ability of traders to identify the best offer available in the market when it is their turn to trade.

Figures 4a and b illustrate the transactions in 100 periods. These figures show that both ZI-U and ZI-C traders trade in MMLOS contexts. The difference between these two figures is that ZI-U buyers trade even at prices that are higher than those they are willing to pay, ZI-C buyers do not. Hence, transactions of ZI-C traders are constrained with the demand schedule.

Simulation time is kept short in order not to crowd these illustrative figures. However, ZI agents continued to trade with each other in experiments as long as 10,000 periods as well and this is not surprising given that there aren't any elements that would stop their trading behaviour in these model settings.

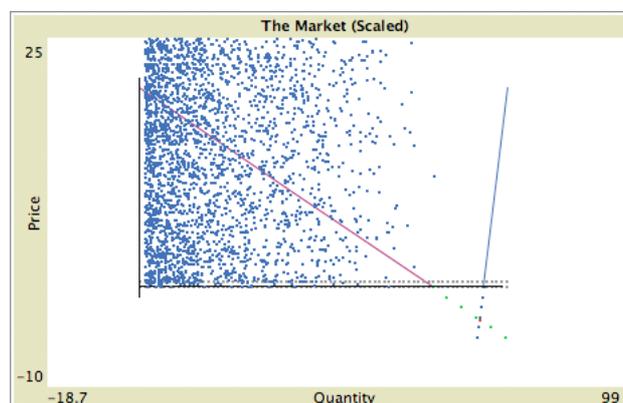


Figure 4a:
MMLOS with ZI-U traders

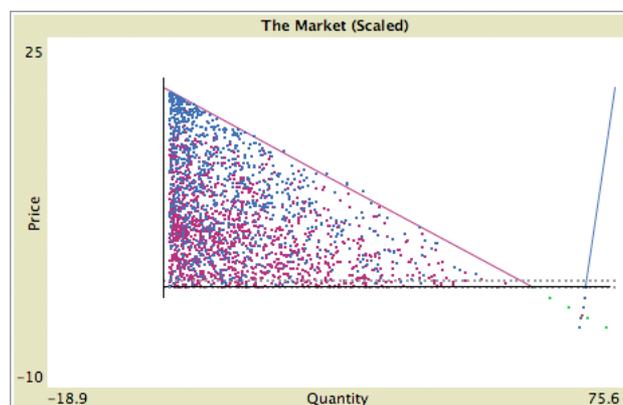


Figure 4b:
MMLOS with ZI-C traders

These results indicate that in the absence of heterogeneous, uncertain and a priori transaction costs, the ZI-C traders would not make a loss, or in the case of buyers, pay more than they are willing to. In the second set of experiments, however, we are allowing for such transaction costs. It is important to highlight that Assumption 2b above describes a sunk cost. Hence, although traders would ideally like to collect these costs back, once they occur, it may be more beneficial for traders under price competition to accept these costs as loss and charge (accept) a lower (higher) price than not to sell (buy) at all.

To represent this property of sunk costs as simply as possible, I equipped the ZI-C traders in the model with a simple learning mechanism. I allowed them to ignore the transaction costs while trading as in the first experiment set, but they now have the simple ability of looking back to previous periods, seeing the price of their last transaction and assessing the profit or buyers' surplus they are likely to get

from their prospective transactions. If this value is lower than their transaction cost in the last period, they lose their interest in trading in this market and drop out.

I run 100 experiments in this second experiment set, and in all of these runs all traders eventually dropped out, leading the market to disappear. In the 100 runs of experiments that I conducted, the simulated time of last transaction varied between 56th and 230th periods with a mean of 118 and standard deviation of 37.3.

Hence, whereas the ZI-U and ZI-C traders who doesn't have any other capabilities of assessing market conditions continue trading in MMLOS conditions, the low intelligence traders who can remember the previous period and make a simple calculation using this memory did not continue to trade keeping everything constant.

This result is not surprising in retrospect. The price competition described above corresponds to a Bertrand game (see, for example, Baye and Morgan, 1999) and since the waste resources are by-products of other productive activities their marginal cost is zero. Since the transaction costs are heterogeneous, uncertain and a priori, the traders cannot coordinate at any positive price level, leaving no incentives for the sellers to trade.

CONCLUSIONS

It is generally taken for granted in the IE literature that in the current organisation of our industrial activities there are mutual economic gains from symbiotic waste resource exchanges that are significant and not adequately exploited. This potential, however, refers to the use-value of waste resources, and not to their exchange value.

This paper examines why some waste resource markets fail to emerge and proposes a simple and well-grounded hypothesis. By relaxing the common assumption of scarcity, the paper illustrates a potential market without a positive equilibrium price. While scarcity and abundance are opposite terms, the lack of scarcity as explained in this paper is not equivalent to abundance. Waste resources are no longer a part of the natural environment and in many cases, cannot be simply put back into nature without causing environmental damage. In addition, in the case of waste resources the good or the resource has a clear owner who needs to be actively involved in order for the exchange to occur. Thus, the story being told here is different from public goods and free goods. Whereas lack of clear ownership and nonappropriation are the common problems for public goods, the very source of the market failure in this case is the difficulty of co-ordination on a positive price and the corresponding lack of incentives to trade.

The main argument of this paper is that in the absence of scarcity, markets may not be able to provide a socially efficient approximation of the use-value. In this respect, the technical possibilities of useful resource exchanges that IE experts identify may not suffice for adequate incentives for economic agents. In the MMLOS case, it is not the technical

properties or the use-value of the waste resource that deters the emergence of the market but expectations about competitors' pricing behaviour.

The MMLOS case is different from and arguably more interesting than Arrow's (1970) missing market for futures contracts. Unlike in Arrow's missing market, it is possible to realise mutually beneficial exchanges in MMLOS, and so, the absence of a market is not strictly Pareto efficient.

The hypothesis is also in compliance with the observed effects of landfill tax on waste resource exchanges, in particular on the appearance of negative prices. The negative price is a relatively new, economically interesting concept and a by-product of a revised conceptualisation of nature that recognises ecosystem services such as pollution absorption as scarce. The peculiarity of negative prices is most clearly exemplified in Baumgartner and Winkler's (2003) study where the price of some waste paper types oscillate between positive and negative prices. With the introduction of landfill tax, negative prices can be more widespread than it is reported in the existing literature.

The paper employs agent-based experiments with zero and low intelligence agents as possibility proofs showing that the lack of scarcity of a waste resource in a given industrial system or region may impede exchanges of this resource and block the formation of corresponding markets. The hypothesis is potentially useful as theoretical background for IE studies and for underpinning the environmental and distributional implications of different business and policy tools.

It is important to clarify that this paper does not suggest that lack of scarcity is the only reason why some waste resource markets may fail to emerge. On the contrary, it is important to stress that this issue is only relevant to incentives to trade. While incentives may be necessary conditions for the emergence of markets, they are not sufficient conditions. This paper defines markets as complex systems of institutions and the emergence of such institutions occurs over time and within a material and social setting.

That lack of scarcity may be the reason why some waste resource markets fail to develop is proposed as a hypothesis which remains to be tested with empirical observations. One of the difficulties of studying missing markets is that it is a negative concept, a case of non-emergence that is difficult to identify and study empirically. In the case of MMLOS, the identification of missing markets can roughly correspond to identification of a waste resource with significant use-value that is not being utilised. It is then necessary to investigate if the lack of scarcity condition is among the reasons of non-emergence.

REFERENCES

- [1] J. R. Anon, M. G. Anon, F. J. Palacios, and L. N. Regueira. Forest waste as a potential alternative energy source. *Journal of Thermal Analysis*, 41:1393–1398, 1994.
- [2] K. Arrow. Political and economic evaluation of social effects and externalities. In J. Margolis, editor, *The Analysis of Public Output*, pages 1–30. UMI, 1970.
- [3] N. M. Asquith, M. T. Vargas, and S. Wunder. Selling two environmental services: In-kind payments for bird habitat and watershed protection in los negros, bolivia. *Ecological Economics*, 65(4):675–684, 2008.

- [4] R. U. Ayres. Sustainability economics: Where do we stand? *Ecological Economics*, 67(2):281–310, 2008.
- [5] N. A. Barr. Economic theory and the welfare state: a survey and interpretation. *Journal of Economic Literature*, 30(2):741–803, 1992.
- [6] F. M. Bator. The anatomy of market failure. *The Quarterly Journal of Economics*, 72(3):351–79, 1958.
- [7] D. Begg, S. Fisher, and R. Dornbush. *Economics*. London: McGraw-Hill, 1994.
- [8] S. Bhattacharya, P. A. Salama, H. Hu Runqingb, H. Somashekarc, D. Racelisd, P. Rathnasirie, and R. Yingyuadf. An assessment of the potential for non-plantation biomass resources in selected asian countries for 2010. *Biomass and Bioenergy*, 29(3):153166, 2005.
- [9] A. C. Carpenter and K. H. Kevin H. Gardner. Use of industrial by-products in urban roadway infrastructure argument for increased industrial ecology. *Journal of Industrial Ecology*, 13(6):965–977, 2009.
- [10] M. R. Chertow and D. R. Lombarti. Quantifying economic and environmental benefits of co-located firms. *Environmental Science and Technology*, 39(17):6535–41, 2005.
- [11] H. E. Daly. The economics of the steady state. *American Economic Review*, 64(2):15–21, 1974.
- [12] S. Dasappa. Potential of biomass energy for electricity generation in sub-saharan africa. *Energy for Sustainable Development*, 15(3):203–213, 2011.
- [13] A. de Janvry, M. Fafchamps, and E. Sadoulet. Peasant household behaviour with missing markets: Some paradoxes explained. *The Economic Journal*, 101(409):1400–17, 1991.
- [14] H. Demsetz. The exchange and enforcement of property rights. *Journal of Law and Economics*, 7(October 1964):11–26, 1964.
- [15] U. S. Department of Energy. *U.S. Billion-Ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry*. 2011.
- [16] D. Gode and S. Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *The Journal of Political Economy*, 101(1):119–37, 1993.
- [17] J. Gressela and A. Zilbersteinb. The forgotten waste biomass; two billion tons for fuel or feed. In F. Johnson and C. Linke-Heep, editors, *Industrial Biotechnology and Biomass Utilisation: Prospects and Challenges for the Developing World*, pages 121–132. SEI and UNIDO, 2007.
- [18] W. P. Heller. Equilibrium market formation causes missing markets. *University of California Discussion Paper*, 93-07R, 1997.
- [19] F. Johnson and C. Linke-Hepp. *Industrial Biotechnology and Biomass Utilisation: Prospects and Challenges for the Developing World*. SEI and UNIDO, 2007.
- [20] N. Johnstone and S. de Tilly. Introduction and overview of market failures and barriers. In *Improving Recycling Markets*, pages 15–50. OECD, 2006.
- [21] J. Kronenberg and R. Winkler. Wasted waste: An evolutionary perspective on industrial by-products. *Ecological Economics*, 68(12):3026–3033, 2009.
- [22] M. D. Kumar and S. O. P. Market instruments for demand management in the face of scarcity and overuse of water in gujarat, western india. *Water Policy*, 3(5):387–403, 2001.
- [23] S. G. Medema. The hesitant hand: Mill, sidgwick, and the evolution of the theory of market failure. *History of Political Economy*, 39(3):331–358, 2007.
- [24] T. Nguyen, J. Slawnwhite, and K. G. Boulama. Power generation from residual industrial heat. *Energy Conversion and Management*, 51(11):2220–2229, 2010.
- [25] A. Randall. The problem of market failure. *Natural Resources Journal*, 23(1):131–148, 1983.
- [26] L. Robbins. *An Essay on the Nature of Significance of Economic Sciences*. London: MacMillan and Co., 1945.
- [27] P. A. Samuelson. *Economics*. McGraw-Hill, 1948.
- [28] R. N. Stavins. What can we learn from the grand policy experiment? lessons from so2 allowance trading. *The Journal of Economic Perspectives*, 12(3):69–88, 1998.
- [29] J. E. Stiglitz and A. Weiss. Credit rationing in markets with imperfect information. *American Economic Review*, 71(3):393–410, 1981.

DEGREE VARIANCE AND EMOTIONAL STRATEGIES CATALYZE COOPERATION IN DYNAMIC SIGNED NETWORKS

Simone Righi
Károly Takács

MTA TK "Lendület" Research Center for Educational and Network Studies (RECENS)
Hungarian Academy of Sciences
Országház utca 30, H-1014, Budapest

Email: simone.righi@tk.mta.hu and takacs.karoly@tk.mta.hu

KEYWORDS

Evolution of cooperation, signed graphs, network dynamics, negative ties, agent-based models, degree heterogeneity

ABSTRACT

We study the problem of the emergence of cooperation in dynamic signed networks where agent strategies coevolve with relational signs and network topology. Running simulations based on an agent-based model, we compare results obtained in a regular lattice initialization with those obtained on a comparable random network initialization. We show that the increased degree heterogeneity at the outset enlarges the parametric conditions in which cooperation survives in the long run. Furthermore, we show how the presence of sign-dependent emotional strategies catalyze the evolution of cooperation with both network topology initializations.

INTRODUCTION AND RELATED LITERATURE

Cooperation among individuals is a key element for the survival and functioning of human and many animal societies. While cooperation is socially optimal, it is difficult to explain its existence in a population of selfish individuals. The Prisoner's Dilemma (PD) is frequently used to study this puzzle as it describes the situation in which the self interest of the individual is opposed to the emergence of cooperation. Two players are given two alternative strategies: to cooperate or to defect. Defection guarantees a higher payoff regardless of what the partner does and is thus the dominant strategy. However, cooperation - if played mutually - provides higher payoffs than mutual defection.

A natural framework in which to study the emergence of cooperative behaviour is evolutionary game theory. This literature burgeoned following the seminar papers of Maynard Smith (1982); Maynard Smith and Price (1973) and Axelrod and Hamilton (1981), with a large number of contributions being dedicated to the puzzle of cooperation (see Rowthorn et al. (2009) for a recent survey of this subject). One strand of this literature looks into the effects of the structure of interactions on the outcome of the evolutionary process. In the context of the single shot PD, they find that cooperation in an unstructured population of randomly

interacting individuals is not viable. Natural selection favors selfish defection, thus leading to groups composed entirely of agents playing this strategy (Hofbauer and Sigmund 1998; Taylor and Jonker 1978). Introducing a more stringent structure for the social contacts and thus allowing only agents that are interconnected on a network to interact (Nakamaru et al. 1997; Nowak and May 1992) seems to provide a solution. Indeed, structuring the interactions increases both realism of models and the realism of conclusions allowing the survival of cooperation in the population. When considering structured interactions, the impact of network topology on the diffusion of cooperation needs can be addressed (Hauert 2004; Johnson et al. 2003; Ohtsuki et al. 2006; Santos and Pacheco 2005) and the realism of model can be improved allowing the interaction structure to co-evolve with agent strategies. In this case, chances for cooperation are enhanced (Santos et al. 2006; Yamagishi and Hayashi 1996; Yamagishi et al. 1994). More specifically, among the mechanisms that improve the conditions of cooperation in dynamic networks are the possibility of partner selection, exclusion of defecting agents, and exit from relationships (Schuessler 1989; Vanberg and Congleton 1992; Yamagishi and Hayashi 1996).

A recent series of our (Righi and Takács 2014a,b) and other authors' papers (Szolnoki et al. 2013), extended the analysis of the emergence of cooperation to signed networks. We introduced the possibility of network ties to turn positive or negative, or to be deleted and relinked as a consequence of previous interactions. We showed that the presence of emotional strategies - that use the *emotional* content implied by the relational signs in social interactions when considering the strategy to play - is pivotal for the survival and diffusion of cooperation. Indeed, in some cases, this strategy acts as a catalyst for unconditional cooperation rather than gaining dominance itself. We characterized the conditions in terms of the speed of evolution and selection pressure that allow the emergence of cooperation. In line with the literature, we found that relatively low rates of strategy adoptions and high rates of rewiring of stressed links are required in order to sustain cooperation.

In this paper, we further extend the study of the emergence of cooperation in signed networks studying the impact of variance (or heterogeneity) in the number of connections of agents

at the outset. In particular, we compare the results obtained on a regular lattice with those obtained on a comparable random network. We show that the increased degree variance at the outset extends the parameters' range under which cooperation survives in the long run. In this sense, our results confirm and extend those of Santos and Pacheco (2005), and show that networks with high heterogeneity in degrees improve the conditions for the emergence of cooperation. Moreover, we show that the benefits in terms of increased space for cooperation by introducing the emotional strategy extend to both random networks and regular lattices.

In the remaining of this paper, we proceed as follows. First we discuss the characteristics of the agent-based model, then we report our results, and conclude with a brief discussion.

MODEL

We consider a population of size N , connected by an undirected and non-weighted signed network. We restrict our interest to networks that are single components. Each agent $i \in \{1, 2, \dots, n\}$ plays the single-shot Prisoner's Dilemma (PD) with each of his current neighbors, i.e. with a subset of the whole population $\mathcal{F}_i^t \subset N$. The cardinality k_i^t of \mathcal{F}_i^t is the degree (or number of network contacts) of the agent i , at time t . The network is signed and each tie is labelled either *negative* or *positive*.

We assume that the social network constrains the possible interactions so that only currently connected agents can play the game together. The payoff structure of the PD is reported in Table 1. When two agents cooperate with each other, each gets a reward (R). When they both defect, they are both punished (P). When one agent defects and the other cooperates, the first gets a temptation payoff (T), while his partner obtains the sucker payoff (S). The PD is defined with payoffs $T > R > P > S$. A typical additional assumption, that we adopt here, is $T + S > R + P$ (Axelrod 1984).

TABLE 1: The Prisoner's Dilemma payoff matrix. The numerical payoffs used here are the same of Axelrod (1984).

	C	D
C	($R = 3, R = 3$)	($S = 0, T = 5$)
D	($T = 5, S = 0$)	($P = 1, P = 1$)

Agents play cooperation or defection in the PD according to their type. We consider three possible strategy types:

- Unconditional Cooperation (UC) that always cooperates, without taking into account the sign of the tie he shares with his interacting partner.
- Unconditional Defection (UD) that always defects.
- Conditional Action (COND) that cooperates with agents he shares a positive tie with and defects with those he has a negative tie with¹. We label this strategy as *emotional*, because it is a trigger response to the valence of the relation.

¹The opposite strategy, that of defecting with cooperators and cooperating with defectors is not considered as deemed to be unrealistic.

We let our agent based model to run in time steps. Steps are iterated until an equilibrium is reached. The conditions for considering one configuration as an equilibrium are stringent. It is required that: (1) a transitory period of 150 steps has passed from the beginning of the simulation (2) in five randomly chosen periods of time since (each time has a probability 0.1 to be selected) the configuration of both relational sign, network topology and agent types needs to be precisely the same.² Each time step (say t), a set of actions are performed by each agent, with the updates being done in parallel. Agents interact with peers they were connected with at the previous time step ($t - 1$) and eventual updates in signs or network topology are observed by partners only in the following step $t + 1$. Following a typical implementation of the literature, we assume that each agent plays the PD with all agents in his first order social neighborhood (i.e. with each $j \in \mathcal{F}_i^{t-1}$) and the average payoff is used when updating the agent strategy. The interested reader can find in Righi and Takács (2014b) a discussion of the effects of using an alternative, sequential, updating protocol.

The dynamics of our model allows for the co-evolution of network signs, agent strategies, and network structure. At each time step, network signs and agents behavior influence each other and the latter also affects the evolution of network topology. More in detail, after each dyadic interaction, stressed network signs are updated (with probabilities P_{neg} and P_{pos}) or deleted and substituted with a new one with a certain probability (P_{rew}). At the end of each time step, when all payoffs are calculated, agents update their strategy to one that has been more successful in their neighborhood, with a certain probability P_{adopt} (see Algorithm 1).

```

for each agent  $i$  do
  Compute its social neighborhood  $\mathcal{F}_i^{t-1} \in N$ ;
  for each agent  $j \in \mathcal{F}_i^{t-1}$  do
    Play the PD and compute payoffs;
    Update the relational sign between  $i$  and  $j$ ;
    If tense, delete the link between  $i$  and  $j$  (with
    Probability  $P_{rew}$ );
  end
  Compute average payoff of agent  $i$ ;
end
for each agent  $i$  do
  Observe the average payoffs of each agent
   $j \in \mathcal{F}_i^{t-1}$ ;
  Adopt the strategy of one agent with (strictly)
  higher payoff (with probability  $P_{adopt}$ );
end

```

Algorithm 1: Intra-step dynamics, repeated at each time step t . Details are provided in the next paragraph.

Let's discuss each of the elements described in Algorithm 1 more in detail.

Update of the relational sign between i and j . After each dyadic PD game, agents might update their relational sign with each other. Given the nature of the PD, there are three possible situations:

²Robustness checks with alternative parametrizations have been performed and they do not influence the results.

- *Both players cooperate.* In this case an existing positive connection remains positive, while a negative one turns positive.
- *Both players defect.* Similarly, an existing positive relation is turned negative and a negative one remains so.
- *One agent cooperates and the other defects.* In this case, the emotional content of the relationship is subject to stress. We assume that if the link is positive, then the cooperator is *frustrated* to have a positive relation to a defector. Therefore, we assume that the valence of the tie can turn negative with probability P_{neg} . If the link is negative, the defector might be interested in turning it into a positive tie. We assume that it happens with probability P_{pos} . There are two possible justifications for such behaviour. The first is that the defector feels remorse or moral guilt (as suggested by Gaudou et al. 2014). The second is instead purely selfish. The defecting partner is content to remain friends with the cooperator. This type of relationship provides him with a strictly higher payoff, in case he is paired with a COND player, whose action is sensitive to the sign of their relationship. It is logical to assume, however, that the frustration from the cooperator is larger, therefore we impose $P_{neg} \gg P_{pos}$.³

Delete the link between i and j and create a new tie. An agent, frustrated by the current behavior of the partner, may decide to delete the social connection completely with probability P_{rew} . In this sense, our network topology co-evolves with agents' strategies endogenously (similarly to Santos et al. 2006). P_{rew} , called rewiring probability, is assumed to be equal for the whole population and it non-strategic. When rewiring takes place, once the old link is erased, a new one is created with another agent. In line with the sociological literature (Granovetter 1973), we assume that there is a tendency towards transitive closure.⁴ New connections are created to friends of friends (excluding the possibility of connections to friends of enemies, to enemies of friends, or to enemies of enemies). In order to introduce some social noise, with a probability P_{rand} , rewiring takes place to a randomly selected agent in the population.⁵ The network structure evolves dynamically through rewiring. This implies that, while the initial topology is either a regular lattice or a random network, it does not necessarily remain of this type - and in general, it does not.

Adopt a better strategy. Agents observe their average payoffs as well as the ones of the agents in their social neighborhood, and are thus able to measure the relative local efficiency of their strategy. If a subset of agents in \mathcal{F}_i^{t-1} has a payoff at time t higher than his own, then agent i will adopt the strategy played in t by one of them, selected uniformly

³For the runs reported here, we fixed $P_{neg} = 0.2$ and $P_{pos} = 0.1$. These values are assumed equal for all agents. We run a sensitivity analysis of this parameter in Righi and Takács (2014a)

⁴The assumption of existence of transitive closure makes the model more realistic and increase cooperation. As shown in Righi and Takács (2014a) however, our results are qualitatively robust when we relax this assumption and consider totally random rewiring.

⁵This parameter is assumed to be small but positive. Its value is fixed to $P_{rand} = 0.01$ in our simulations.

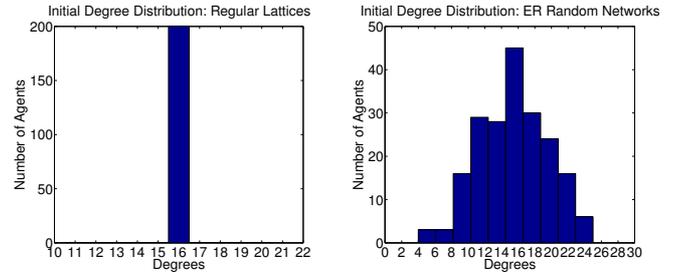


Fig. 1: Degree distributions in typical networks used for initialize our simulations. In both the regular lattice case (Left Panel) and the random network case (Right Panel) the average number of connections per agent is 16. $N = 200$.

at random. Evolutionary update happens, for each agent, with probability P_{adopt} which is assumed to be equal for all agents.

Simulations Calibration. Concerning the initial structure of the social network, we provide results for two cases. In a first set of simulations we assume that agents are laid on a *regular lattice* in which every agent has precisely 16 connections (the degree distribution is therefore degenerate as shown in the Left Panel of Figure 1). Then we introduce heterogeneity in degree distribution and we study networks initialized as Erdős-Rényi (Erdős and Rényi 1959) *random graph* (an example of the resulting degree distribution is provided in the Right Panel of Figure 1). In order to make the results comparable we impose that each pair of nodes is connected with an independent probability $P_{link} = 0.16$ so that the degree distribution is centered around 16 with a standard deviation of about 4. Moreover, agents are assigned with one of the three strategies randomly in equal proportions. In the absence of conditional players, the proportion of UDs and of UCs are 1/2. When CONDs are added, then the starting proportion of each type of agent is 1/3. Finally, the relational signs are randomly distributed and initialized so that each link has a 50

RESULTS

Our aim with this study is to characterize the parameter configurations that favor the evolution of cooperation in dynamic signed networks. We focus on the effect of conditional (or emotional) strategy in two different network initializations: in a regular lattice and in an Erdős-Rényi random graph. The two main dynamic forces that operate in our model are the evolutionary pressure (P_{adopt}) and the network update dynamics (P_{rew}). Our strategy is to analyze their impact, changing their relative strength progressively. For each possible combination of the two probabilities (each studied for values between 0 and 1 with a granularity of 0.05) we show results concerning the average proportion (calculated in 50 simulations) of the agents and network ties surviving at the steady state.⁶

Figures 2 and 3 report results for two alternative cases each. In Figure 2, the social network is initialized so that every

⁶Standard deviations are not reported here and are available upon request. The variability of the results is quite small except in the area of the phase transition between the configurations in which cooperation survives and those where it disappears completely. Only statistically significant phenomena are studied and discussed in the following.

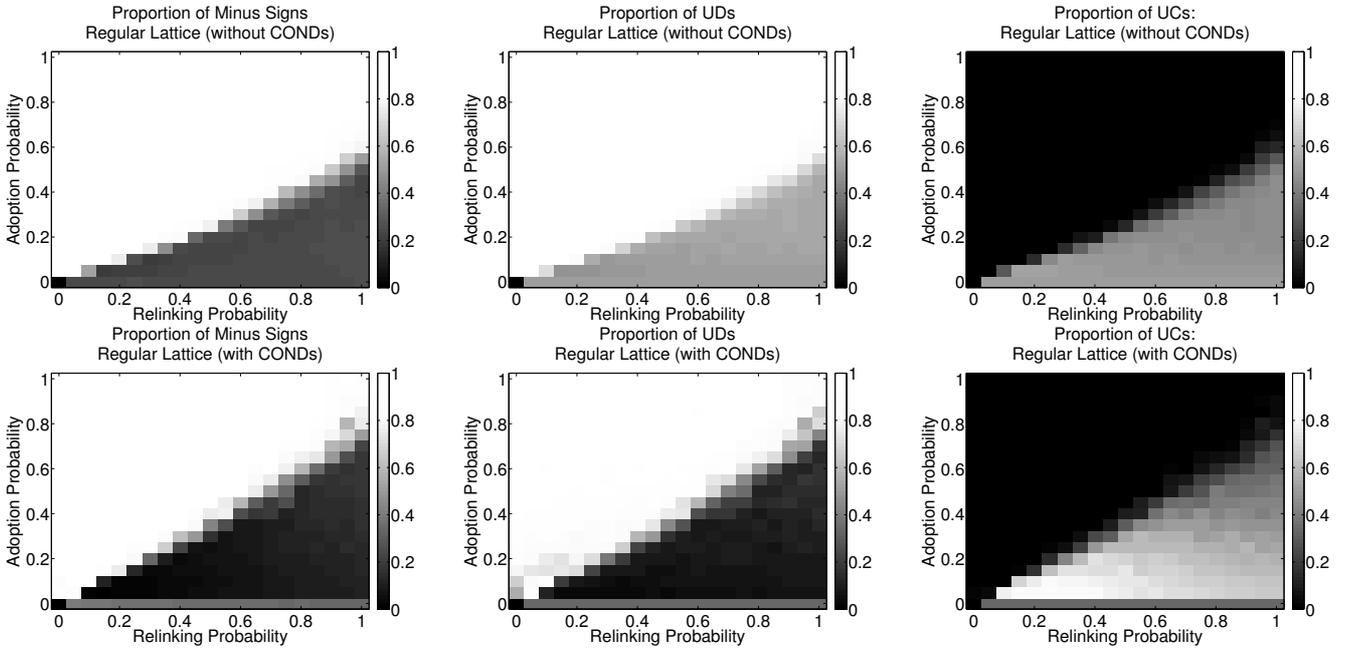


Fig. 2: Effect of the competing dynamics of strategy adoption (vertical axis) and of rewiring of stressed links (horizontal axis) on the final proportion of negative ties in the network (Left Panels), of UD (Central Panels) and of UCs (Right Panels). Top Panels show results for populations initialized as equally divided between UCs and UD (where there are no CONDs). Lower Panels show results for populations initialized as equally divided among the three different agent types. The social networks are initialized as *regular lattices* where each agent has 16 connections. Each datapoint is the average of 50 simulations. For each simulations $N = 200$ and network signs are randomly initialized with equal probability.

agent has initially the same amount of connections, which defines a regular lattice. In Figure 3), results are shown for a setup where degree variance is introduced and the network is initialized as an Erdős-Rényi random network. In the Top Panels of the Figures, the population is initialized as equally divided between unconditional cooperators and unconditional defectors. This simulation is compared with the case (Lower Panels), in which the population is initialized as divided equally among UCs, UD and CONDs.

One can observe several similarities in the results from the two kinds of starting configurations. As noted in our previous contributions (Righi and Takács 2014a,b), and coherently to what observed by Santos et al. 2006, the relative speed (i.e. the probability) of the network topology update and of strategy adoption have two opposite effects on the viability of cooperative strategies. Increasing the speed of adoption of better strategies favors defection, as this is the strategy that maximizes payoffs in dyadic terms. At the opposite, a relatively high degree of network updating leads to higher proportions of cooperation, as it helps the formation of clusters of cooperators.

From Figures 2 and 3, we can observe that defectors suddenly lose dominance when a certain ratio between the two dynamic forces is reached. In the case of the regular lattice initialization without emotional strategies, the cooperation survives if the approximate relation $P_{rew} > 2P_{adopt}$ holds. The chances of cooperation are increased for ER networks compared to a regular lattice initialization for any combination of P_{rew} and P_{adopt} . In this case, the condition for cooperation

to survive is $P_{rew} > 5/3P_{adopt} (\sim 1.6P_{adopt})$.

Let's speculate about the reason for this improvement. In the absence of CONDs, the only force preventing UCs from being eliminated from the network is the rewiring of stressed ties. As we discussed, this process tends to create clusters of cooperators that can then survive. When all agents have the same connectivity, they all require similar amounts of time to rewire their connections to UD, which can then spread locally and dominate the final population. When the network is initialized as random, some of the agents have less connections than the average, and they become isolated from the defector faster; substituting negative stressed connections with positive ones via transitive closure. The new connections are more likely to be with CONDs (when present) or UCs (which cooperate at least when given the opportunity and thus tend to develop positive ties) than with defectors, given the positive relations involved. These agents constitute therefore the nucleus of cooperative clusters around which more connected cooperative peers can survive.

In summary, degree heterogeneity provides *time* for clusters to form, even when the ratio between P_{rew} and P_{adopt} is less favorable. The positive effect of the increased heterogeneity for cooperation is stronger in more dynamic networks (higher P_{rew}) since agents with few connections extricate more efficiently their leverage effect on the formation of cooperative clusters. Introducing a variability in the degrees of the agents, thus increases the range of parameters in which cooperation survives and diffuses in the population. This result confirms the one obtained by Santos and Pacheco (2005). We consider,

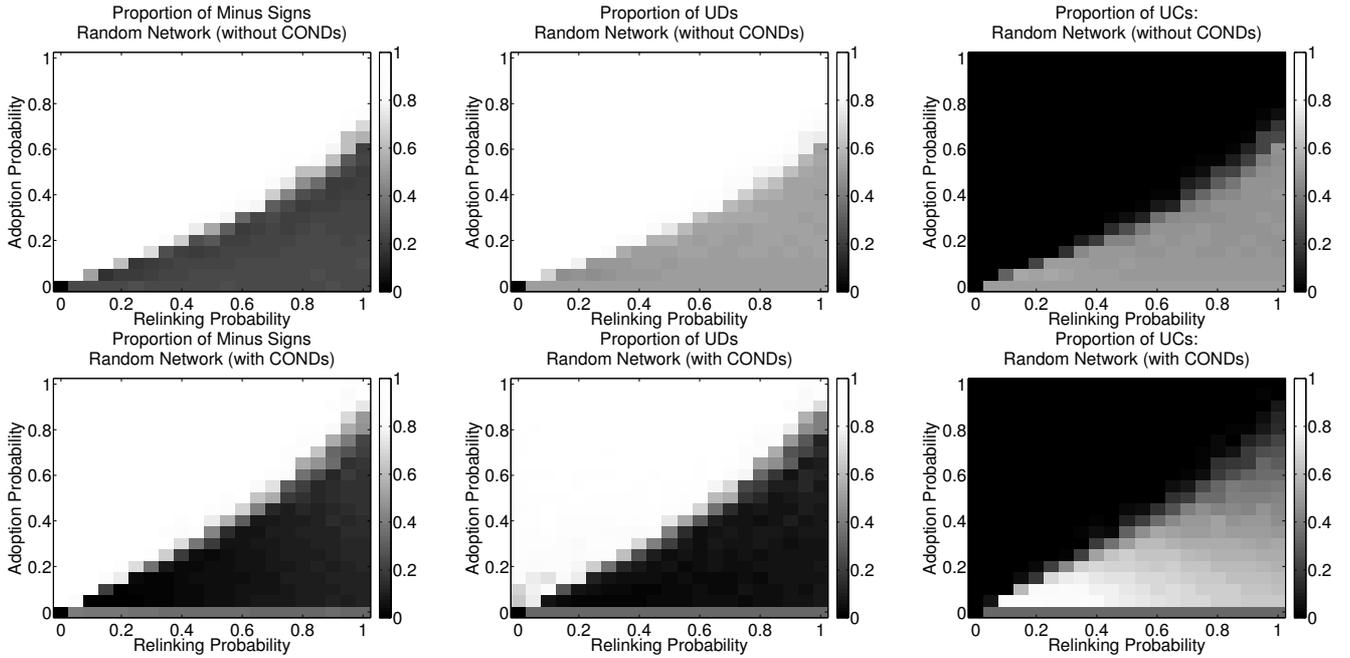


Fig. 3: Effect of the competing dynamics of strategy adoption (vertical axis) and of rewiring of stressed links (horizontal axis) on the final proportion of negative ties in the network (Left Panels), of UDs (Central Panels) and of UCs (Right Panels). Top Panels show results for populations initialized as equally divided between UCs and UDs (where there are no CONDs). Lower Panels show results for populations initialized as equally divided among the three different types of agents. The social network are initialized as Erdős-Rényi *random network* with $P_{Link} = 0.16$. Each datapoint is the average of 50 simulations. Network signs are randomly initialized with equal probability. $N = 200$.

however, a dynamic environment in which agent strategies co-evolve with relational signs and network topology. Moreover, from the purely topological point of view, we show that cooperation can be increased through heterogeneity also without recurring alterations of the randomness of the network (such as preferential attachment or network growth).

In both types of network initializations, when the conditions for cooperation to survive are met and CONDs are absent, the results for different parameter combinations are rather similar. They indicate that about 25% of the signs in the network are negative and the population turns out to be equally split between UCs and UDs. Regardless of the starting network, we observe that when cooperation survives, it does not diffuses. In both cases, the proportion of cooperators remains similar (or just below) its initial setup value. We can understand this result observing that the sole driving force allowing the survival of cooperation (in the absence of emotional strategies) is the rewiring of stressed links. In this sense, P_{rew} has a purely positive effect on cooperation and P_{adopt} has a purely negative one. When the first dominates, cooperation survives, when the second dominates cooperation disappears; hence there is the sharp phase transition between the two states.

As noted in our previous work (Righi and Takács 2014a), the introduction of the COND strategy relaxes the parametric conditions for cooperation to survive. In the context of this paper, we note that this happens both for the regular lattice initialization, where the approximate condition for the survival of cooperation becomes $P_{rew} > 20/13P_{adopt} (\sim 1.53P_{adopt})$,

and for the random network where it becomes $P_{rew} > 4/3P_{adopt} (\sim 1.3P_{adopt})$. The relative effect of introducing CONDs in the population is thus similar in terms of the proportion of parameters in which cooperation becomes viable and thus the two initializations can be discussed together.

Understanding this result requires a closer look at the final proportions of agents and relational signs in the area that allows the survival of cooperation. Here, the final proportion of UC agents ranges from 25% to 75% of the population, with this probability decreasing monotonically as the adoption and rewiring probability increases. Confronting the results regarding UCs with those regarding UDs, one can notice that the decrease in the proportion of UCs benefits UDs little (their proportion passes from a minimum of about 5% to a maximum of 17%, but much more the conditional players (see Figure 4). In the same area, also the proportion of negative ties progressively increases from about 4% to about 25%, but never exceeds this value. We can thus conclude that, when cooperation is viable, clusters are formed in which CONDs and UCs are intertwined by positive links and therefore are functionally indistinguishable. Moreover, while in presence of conditional agents cooperative behaviour spread in the population, the dominant type of cooperation (conditional or unconditional) depends on the relative strength of our two main dynamic variables.

For environments with relatively *low frequency* of network and evolutionary updates, the COND strategy acts as a shield for UCs. When in contact with both pure cooperators and pure defectors, agents playing this strategy tend to enjoy higher

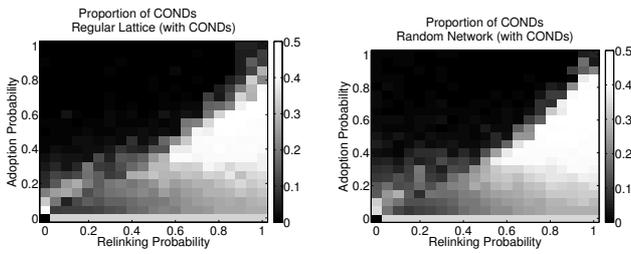


Fig. 4: Effect of the competing dynamics of strategy adoption (vertical axis) and of rewiring of stressed links (horizontal axis) on the final proportion conditional (COND) agents. Left Panel: results for the regular lattice initialization. Right Panel: results for the random network initialization

payoffs than those who always defect (COND gets payoffs from cooperation when interacting with the UC type, while avoiding the *sucker* position when interacting with UD), and therefore tends to replace them due to the evolutive pressure. As the proportion of UDs decreases and segregation increases, pure cooperation becomes the optimal strategy, as it avoids "errors" due to the mis-interpretation of network signs. Thus, UCs tend to diffuse at the expense of CONDs, and the final proportion of unconditional cooperators tends to be high. At the opposite, when the two dynamic updates happen relatively fast, this dynamics reverse in favor of COND. When adoption of strategies with higher payoffs is faster, the cooperation is in general more difficult to sustain and pure defection diffuses more. Under these conditions, emotional agents, being able to discern among cooperators (with whom they tend to form positive ties) and defectors (with whom they tend to form negative ties) suffer less *sucker* payoffs from pure defection than pure cooperators; which therefore tend to disappear faster. In this more dynamic setup, the number of cooperators reduces too fast to regain dominance later, and conditional cooperation turns out as dominant.

CONCLUSIONS

The problem of evolution of cooperation has been widely studied in social sciences. Unlike most of the previous literature that considered only positive relations, we introduced negative ties as a force that is able to influence agents behavior. We presented results from an agent based model, where we studied the evolution of cooperation in dynamic signed networks in which agent strategies co-evolved with relational signs and network topology. Agents played the Prisoner's Dilemma with their current neighbors and the result of dyadic interactions drove the evolution of relational signs and network relations. The average performance of a strategy across all interactions of one individual was defined as the fitness value that determined the evolutionary process.

In this paper, we performed an extensive simulation analysis of our model focusing on the effects on the survival of cooperative strategies as (1) network topology was varied, considering a regular lattice and a random network initialization; and as (2) a sign dependent strategy that considers the network signs when deciding whether to cooperate or to defect was introduced. We provided results for all possible combinations of the two main dynamic forces of this model

by progressively changing both the probability of adopting more fitting strategies and the probability of rewiring tense connections.

In all cases, and in line with the literature, we showed that higher strategy evolution rates reduced the combinations of parameters in which cooperation survived (favoring the dyadic dominant strategy: to defect), while increasing the rewiring probability helped isolating cooperators from defectors thus favoring the survival of cooperation.

Random networks provided more place for the emergence of cooperative behavior for a larger set of parameters than regular lattices. This result is similar to the one of Santos and Pacheco (2005), however our outcome follows from a different mechanism. In Santos and Pacheco (2005), cooperation diffusion followed from the presence of very connected hubs. In our setup, there were no such hubs (degrees have a bell shaped distribution around a characteristic degree and connections are purely random). The process allowing the diffusion of cooperation is the presence of individuals which are less connected than the average. By segregating early on in the simulations from defectors, these created the nuclei around which cooperative clusters could emerge.

Extending the analysis of Righi and Takács (2014a) to regular lattices, we studied the effects of the introduction of a conditional strategy that considers the relational signs to the partner to decide whether to cooperate or defect. The conditional strategy enlarged the parametric space in which cooperation evolved. Despite the advantage of being able to use more information, however, and regardless of the network topology adopted, the conditional strategy gained dominance itself only in a few, rapidly changing environment (where both adoption and rewiring happened relatively frequently). In these situations, the better performance of the COND strategy against pure defectors made the spread in the population possible. When the network and strategies were more stable, the conditional strategy acted instead as catalyst for the diffusion of unconditionally cooperative behavior.

The work presented in this paper is a first step in understanding the role of network topology in the diffusion of cooperation in dynamic signed networks. While a preliminary analysis has been introduced in Righi and Takács (2013), further studies are required to address the issue of the evolution of cooperation in non-random signed networks systematically. In particular, more realistic network initializations, such as scale-free and small-world networks, could be analyzed.

ACKNOWLEDGMENTS

The authors wish to thank the "Lendület" program of the Hungarian Academy of Sciences and the Centre for Social Sciences for their financial and organizational support and three anonymous referees for their useful comments.

AUTHORS BIOGRAPHIES

SIMONE RIGHI is currently a Research Fellow at "Lendület" Research Center for Educational and Network Studies (RECENS) of the Hungarian Academy of Sciences. He studied mathematics and economics at the University of Namur (Belgium) where he obtained a Ph.D. in Economics in

2012 with a thesis on "Information aggregation and Political Economics". His main research interests are: opinion dynamics, evolutionary game theory, network theory and industrial organization. His e-mail address is: simone.righi@tk.mta.hu and up to date information and research papers can be found at his web-page: <http://perso.fundp.ac.be/~srighi/>.

KÁROLY TAKÁCS is Director of the "Lendület" Research Center for Educational and Network Studies (RECENS) of the Hungarian Academy of Sciences and associate professor of Sociology at the Corvinus University of Budapest. He obtained his Ph.D. in Sociology from the University of Groningen (Netherlands) with a thesis on "Social Networks and Intergroup Conflict". During and after his graduate studies he has been visiting scholar at Cornell University, University of Brescia and at the Netherlands Institute for Advanced Study in Humanities and Social Sciences. His main research interests are: social networks, intergroup conflict, discrimination, evolution of altruism and of cooperation, agent-based simulations and experiments. His e-mail address is: takacs.karoly@tk.mta.hu. Up to date information and research papers can be found at his web-page: <http://web.uni-corvinus.hu/~tkaroly/>.

REFERENCES

- Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.
- Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.
- Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297.
- Gaudou, B., Lorini, E., and Mayor, E. (2014). Moral guilt: An agent-based model analysis. In *Advances in Social Simulation*, pages 95–106. Springer.
- Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, pages 1360–1380.
- Hauert, C. (2004). Virtuallabs. <http://www.univie.ac.at/virtuallabs/Moran/>. Accessed: 2010-09-30.
- Hofbauer, J. and Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.
- Johnson, J. C., Boster, J. S., and Palinkas, L. A. (2003). Social roles and the evolution of networks in extreme and isolated environments. *Journal of Mathematical Sociology*, 27(2-3):89–121.
- Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge university press.
- Maynard Smith, J. and Price, G. (1973). The logic of animal conflict. *Nature*, 246:15.
- Nakamaru, M., Matsuda, H., and Iwasa, Y. (1997). The evolution of cooperation in a lattice-structured population. *Journal of Theoretical Biology*, 184(1):65–81.
- Nowak, M. A. and May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, 359(6398):826–829.
- Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505.
- Righi, S. and Takács, K. (2013). Signed networks, triadic interactions and the evolution of cooperation. *arXiv preprint arXiv:1309.7698*.
- Righi, S. and Takács, K. (2014a). Emotional strategies as catalysts of cooperation in signed networks. *arXiv preprint arXiv:1401.1996*.
- Righi, S. and Takács, K. (2014b). Parallel versus sequential update and the evolution of cooperation with the assistance of emotional strategies. *arXiv preprint arXiv:1401.4672*.
- Rowthorn, R. E., Guzmán, R. A., and Rodríguez-Sickert, C. (2009). Theories of the evolution of cooperative behaviour: A critical survey plus some new results. *MPRA Paper No. 12574*.
- Santos, F. C. and Pacheco, J. M. (2005). Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9):098104.
- Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Cooperation prevails when individuals adjust their social ties. *PLoS Computational Biology*, 2(10):e140.
- Schuessler, R. (1989). Exit threats and cooperation under anonymity. *Journal of Conflict Resolution*, 33(4):728–749.
- Szolnoki, A., Xie, N.-G., Ye, Y., and Perc, M. (2013). Evolution of emotions on networks leads to the evolution of cooperation in social dilemmas. *Physical Review E*, 87(4):042805.
- Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1):145–156.
- Vanberg, V. J. and Congleton, R. D. (1992). Rationality, morality, and exit. *The American Political Science Review*, pages 418–431.
- Yamagishi, T. and Hayashi, N. (1996). Selective play: Social embeddedness of social dilemmas. In *Frontiers in social dilemmas research*, pages 363–384. Springer.
- Yamagishi, T., Hayashi, N., and Jin, N. (1994). Prisoner's dilemma networks: selection strategy versus action strategy. In *Social dilemmas and cooperation*, pages 233–250. Springer.

A SOCIAL INTERACTION MODEL FOR CRIME HOT SPOTS

Evan C Haskell

Division of Math, Science, and Technology
Farquhar College of Arts and Sciences
Nova Southeastern University
Fort Lauderdale, FL 33314 USA
haskell@nova.edu

KEYWORDS

environmental criminology, social interaction modeling, opinion dynamics, crime hot spots, non-local aggregation

ABSTRACT

A common feature of mapped crime patterns is a strong spatial and temporal clustering into crime “hot spots”. In this paper we explore a social interaction model for the evolution of the attractiveness of the crime environment for criminal activity. We see how hot spots may arise when the idiosyncratic attractiveness of the environment is not encouraging for criminal activity. The stability of these hot spots is determined to depend on both the size of the hot spot and the social interaction function itself.

INTRODUCTION

Criminal activities ranging from homicides to burglaries are unevenly distributed within an environment and amongst victims and offenders (Johnson et al., 2007; Johnson, 2010). The strong clustering of elevated levels of crime in space and time is often referred to as a crime “hot spot”. The environment in which the crime occurs may play a role in the generation and accessibility of crime opportunities and may even provoke criminal activity. Studies of the influence of the role of these opportunities in crime occurrence date back to the nineteenth century (Weisburd et al., 2009; Johnson, 2010); yet, systematic studies of the interaction between the offender and the environment is a relatively recent pursuit in what has been coined environmental criminology (Brantingham and Brantingham, 1981).

While much focus has been placed on the mapping of crime hot spots, and sociological theories developed to account for potential causes of the formation of crime hot spots; mathematical modeling to aid in understanding the mechanisms of the genesis, spread, and dissipation of crime hot spots is in the nascent phase. Mechanistic models have explored agent based simulation and reaction diffusion approaches to the risk of victimization (Short et al., 2008, 2010), social interaction models of the propensity of the criminal agent to act (Berestycki and Nadal, 2010), and incorporating Levy flights to describe non-local movement of offenders (Chatu-

rapruek et al., 2013).

We develop and explore a model for the nonlocal aggregation of environmental attractiveness for criminal activity. The model brings together the ideas of routine activity theory, crime pattern theory, and rational choice theory to explore how social interactions amongst recent and potential victims and offenders influence the aggregation of environmental attractiveness for criminal activities and the formation of crime hot spots. Routine activity theory asserts that societal organization from the routine activities of victims to placement of ‘guardians’ impacts the attractiveness of the victim for victimization (Felson, 2008). Crime pattern theory asserts that the spatial organization of crime concentration reflects the collective awareness of offenders to suitable crime opportunities (Brantingham and Brantingham, 2008). Rational choice theory asserts that the decision to continue or desist from criminal activity by the offender is based on an assessment of the relative risks and potential rewards of the criminal act as perceived by the offender (Cornish and Clarke, 2008). The model uses these theories to present a realization of the ‘broken windows theory’ that signs of disorder attract more disorder and diminishing those signs will diminish the attraction of disorder (Keizer et al., 2008).

The model presented is based on considerations taken in the Berestycki and Nadal model (Berestycki and Nadal, 2010); however, we perform an analysis of the equilibrium solutions when an opinion dynamic for criminal activity is formed through a social interaction of environmental factors that influence the continuance or desistance of crime at a location. The resultant model that we analyze is a reformulation of the seminal model of Amari for the formation of localized activity states in lateral inhibition type neural fields (Amari, 1977). We follow the analysis of Amari to present the necessary and sufficient conditions for the potential equilibrium solutions including hot spots and provide the corresponding taxonomy of equilibrium solutions based upon an idiosyncratic attractiveness of the environment to crime. In the Amari analysis the stability of equilibrium solutions is exhibited through considerations of the stability of the width of the equilibrium solution. Departing from Amari’s analysis, we will present the stability analysis by a standard linearization technique of the

model system. The mechanisms we incorporate in this model are generic mechanisms that could apply to many different criminal activities and we do not specialize to any one crime type. However, the model should be considered as describing the attractiveness for one crime type as the description of social interaction given here is representative of a ‘communication of risk’ about areas where crime has occurred.

MODEL

The object of interest in developing maps of crime hot spots is the level of criminal activity, $u(x, t)$, at some position x in the domain Ω and time t . We assume the level of criminal activity to depend on the typical perceived attractiveness of a given location, x , in the environment for criminal activity, $A(x, t)$. That is, $A(x, t)$ represents a coarse grained view of the environment that can be thought as describing a typical reward ($A > 0$) or risk ($A < 0$) for committing a criminal act at location x on time t . In line with routine activity theory, the dynamics of this attractiveness will depend on the presence of potential offenders, targets, and deterrent forces (Cohen and Felson, 1979; Felson, 2008; Berestycki and Nadal, 2010) This attractiveness variable is analogous to the risk of victimization modeled by Short *et. al.* (Short et al., 2008, 2010) and is an alternative interpretation of the propensity to act modeled by Berestycki and Nadal (Berestycki and Nadal, 2010). The spatio-temporal field $A(x, t)$ may represent, for example, general environmental cues or specific offender knowledge about the vulnerability of the area for criminal activity (Short et al., 2010).

The crime level is considered to be a non-linear function of the attractiveness of the environment

$$u(x, t) = \Lambda[A(x, t)] \quad (1)$$

where the ‘acting out’ function $\Lambda[A]$ is a monotonically non-decreasing, saturating function satisfying $\Lambda[A] = 0$ for $A \leq 0$. That is, for relatively unattractive environments there is no crime and as attractiveness increases so will the crime level, approaching some maximal crime level normalized to a value of 1. For simplicity in the proceeding analysis we will consider the acting out function to be a step function

$$\Lambda[A] = \begin{cases} 0 & \text{if } A \leq 0 \\ 1 & \text{if } A > 0 \end{cases} \quad (2)$$

This choice reflects the binary nature of the decision to act out, or, perform the criminal activity.

We model the time evolution of the attractiveness of the environment incorporating an opinion dynamic represented as a social interaction term similar to that presented by Berestycki and Nadal (Berestycki and Nadal, 2010):

$$\tau \frac{\partial A}{\partial t} = -A(x, t) + W(x, t) + \int_{\Omega} j(x, x') u(x', t) dx' \quad (3)$$

$W(x, t)$ describes the inherent level of attractiveness for criminal activity at a location x and t in the absence of criminal activity. While the inherent attractiveness of a location could be modified over time, we consider this timescale to be long relative to the dynamics of hot spot formation and consider $W(x, t)$ to be time independent. Furthermore, we will make the simplifying assumption that the environment we are considering is uniform in the inherent attractiveness; that is, $W(x, t) = w$ where w is the average value of $W(x, t)$ over the spatial domain. We consider the field to be homogeneous; that is, the weighting function depends only on the distance between locations x and x' and not the specific locations in the environment; that is, $j(x, x') = j(|x - x'|)$. Furthermore, the crime hot spots that we consider are stable persistent elevations of the crime level which is tantamount to an equilibrium solution of (3). Therefore in this description of the social interaction we neglect any time lag between crime events reflected in the crime level and the corresponding impact on the attractiveness of the environment. Both the temporal and spatial scale for the dynamics of attractiveness represented by τ and description of $j(|x - x'|)$ respectively, are not clear from available data. As such we rescale our time units to be in terms of the attractiveness time scale, $t \rightarrow \frac{t}{\tau}$. With this normalization τ is set to unity in equation (3). Additionally, spatial location x is given in terms of the attractiveness spatial scale.

Knox analysis of data for various types of crimes in numerous locations demonstrate an ubiquitous feature of co-occurrences of criminal events that are proximal in time and space that are significantly more common than would be expected if the occurrence of criminal activities were random events (Johnson et al., 2007; Johnson, 2010). This near repeat victimization may reflect a foraging strategy where offenders utilize knowledge from previous activities to assist in future targeting decisions (Johnson, 2010). Thus there is a communication of risk that is reflected by the weighing function for social interaction $j(|x - x'|)$. The Knox analysis suggest that the strengthening of the attractiveness for future crime events is strongest at the same location as where an event has occurred and decreases as the proximity decreases. Just as knowledge of routine activities of victims in an area may increase the attractiveness of the area, so would knowledge about deterrents against criminal activities in an area decrease the attractiveness of the area. Given the information from the Knox analysis it is reasonable to assert that in areas where crime has occurred these deterrents are not as strong as the factors that would increase the attractiveness; however, at a sufficient distance from an area where crime has occurred elevations in co-occurrence of criminal activities diminish suggesting that the extent of deterrent forces to the attractiveness of an area is broader than those that would

enhance the attractiveness for a criminal activity. In fact, the presence of a guardian or deterrent force in response to elevated crime levels does not necessarily displace crime to an adjacent setting (Keizer et al., 2008; Short et al., 2010) indicating a potential greater distal impact of the deterrent force relative to an attractive force. Alternatively, the broader extent could represent an optimal foraging behavior whereby offenders concentrate towards areas of high attractiveness. The opportunity for crime to occur requires the presence of both victims and offenders; hence, an absence of offenders creates a decreased attractiveness for crime to occur.

A social interaction function, $j(|x|)$, that encompasses near-repeat victimization via foraging behavior of criminals and the same mechanisms for the spread of deterrent information through the environment should have a positive local maxima at $|x| = 0$, one root preceding a local minima which is negative, and $\lim_{|x| \rightarrow \infty} j(|x|) = 0$. An example of such a function is given by the difference of Gaussians function

$$j(x) = \frac{j_1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) - \frac{j_2}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$

where $j_1 > j_2 > 0$ and $\sigma_2 > \sigma_1 > 0$.

Previous work (Berestycki and Nadal, 2010) considers a social interaction weighting function of the form

$$j(x, x') = j_0 + j_1(x, x') \quad (4)$$

where $j_1(x, x')$ is positive for ‘‘close’’ locations x and x' and zero otherwise. Under such assumptions the integral term in (3) is well approximated by a diffusive form:

$$\int_{\Omega} j(x, x')u(x', t)dx' \approx j_0\bar{u}(t) + D\nabla^2u(x, t)$$

where \bar{u} is the average crime level over the domain. Additionally in that work social deterrence is incorporated as a separate field that modulates an effective cost to the propensity to commit a crime at a location where there is a non-zero crime level. Separate analysis are provided in that work for the case where there is no social interaction term and deterrence is purely local (i.e. no influence on deterrence level at position x from any other location in the environment) and the case where there is no deterrence and social interaction is global providing an equal influence at all locations (i.e. $j_1(x, x') \equiv 0$ in (4)).

In the subsequent analysis we will consider the domain $\Omega = \mathbb{R}^1$. With this domain, the model presented here is a reinterpretation of Amari’s seminal model for describing neural fields with lateral inhibition (Amari, 1977) to describing crime dynamics. We use techniques developed for analyzing this neural field model and reformulate the results for the given application in crime dynamics.

EQUILIBRIUM SOLUTIONS

Hot spots are characterized as relatively stable localized areas where persistent criminal activity is concentrated. As such to determine necessary and sufficient conditions for the existence and stability of hot spots we analyze equilibrium solutions of (3). At equilibrium $\frac{\partial A}{\partial t} = 0$ and corresponding equilibrium solutions satisfy

$$A(x) = w + \int_{\Omega} j(|x - x'|)u(x')dx'. \quad (5)$$

From the description of the ‘acting out’ function (1), $u(x) = 0$ for $A(x) \leq 0$ so we define the opportunity sets for criminal activity from the set function:

$$R[A] = \{x|A(x) > 0\}$$

which is the region of the field that is considered attractive for a criminal act and hence has a non-negative crime level. With this notation we rewrite (5) as

$$A(x) = w + \int_{R[A]} j(|x - x'|)dx'. \quad (6)$$

Any equilibrium solution with $R[A] = \emptyset$ (i.e. $A(x) \leq 0$ for any x) will be termed a ‘quiet’ or \emptyset -spot. Any equilibrium solution with $R[A] = (-\infty, \infty)$ (i.e. entire field is attractive to criminal activity) will be termed a ‘rampant’ or ∞ -spot. A hot-spot is a localized region of elevated (non-zero) crime level which is represented here as a finite interval for which $A(x) > 0$. that is, $R[A] = (h_1, h_2)$. Given the homogeneity of the field, without loss of generality we can consider a hot spot of length h to satisfy $R[A] = (0, h)$ and will refer to such equilibrium solutions as h -spots.

Before identifying the necessary and sufficient conditions for the existence of \emptyset -, ∞ -, and h -spot solutions we define some pertinent features of the integral term in equation (6),

$$J(x) = \int_0^x j(x')dx'.$$

From this definition, $J(0) = 0$ and $J(-x) = -J(x)$. We further define the quantities:

$$J_{\infty} = \lim_{x \rightarrow \infty} J(x)$$

$$J_m = \max_{x > 0} J(x).$$

Theorem 1: Necessary and Sufficient conditions for equilibrium solutions.

- There exists a \emptyset -spot if and only if $w < 0$.
- There exists a ∞ -spot if and only if $2J_{\infty} > -w$.
- There exists a h -spot if and only if $w < 0$ and $h > 0$ satisfies $w + J(h) = 0$.

Proof:

- If there exists a \emptyset -spot then $A(x) = w$ and $R[A] =$

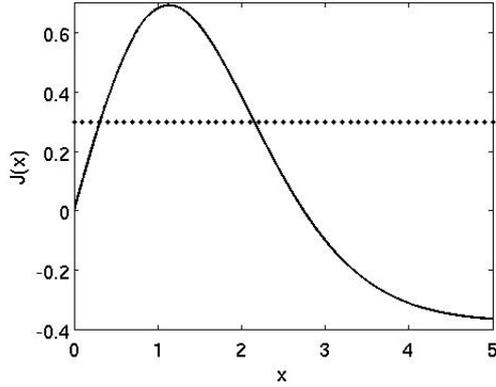


Fig. 1. Finding h -spot widths from $J(x)$: h -spots are found from the intersection of the graphs for $J(x)$ (solid line) and $-w$ (dotted line). In this example we find two h -spots with widths $h_1 < h_2$. Parameters: $j_1 = 7.5, \sigma_1 = 1, j_2 = 8.3, \sigma_2 = 1.65$

\emptyset requires $w < 0$. Conversely, if $w < 0$ then $A(x) = w$ is a \emptyset -spot solution.

b) If there exists a ∞ -spot then $A(x) = w + \int_{-\infty}^{\infty} j(x - x') dx' = w + 2J_{\infty} > 0$. Conversely, if $2J_{\infty} > -w$ then $A(x) = w + 2J_{\infty}$ is a ∞ -spot solution

c) If there exists a h -spot then

$$\begin{aligned} A(x) &= w + \int_0^h j(x - x') dx' \\ &= w + J(x) - J(x - h). \end{aligned} \quad (7)$$

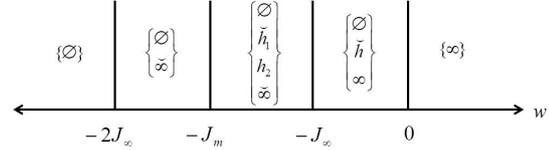
From the continuity of $A(x)$, $A(0) = A(h) = 0$ implying $w + J(h) = 0$. Conversely if $w + J(h) = 0$ then a solution defined by (7) satisfies $A(0) = A(h) = 0$. Furthermore $\frac{dA}{dx} = j(x) - j(x - h)$ which is positive at $x = 0$ and negative at $x = h$ so $A(x) > 0$ for $0 < x < h$ and provided w is sufficiently negative, $A(x) < 0$ outside this interval. \square

From the theorem 1 we note that the existence of the various types of equilibrium solutions depend on the interaction of the idiosyncratic attractiveness of the environment and the properties of the social interaction function. For example, to have a zero crime level requires the idiosyncratic attractiveness of the environment to reflect an expected risk to engaging in criminal activity. However this alone does not guarantee a zero crime level environment, if the net social interaction is large enough relative to the idiosyncratic attractiveness, then there is potential for rampant crime where there is a non-zero crime level throughout the environment. Theorem 1 allows us for a given social interaction function, $j(|x|)$, to develop a taxonomy of hot spot equilibrium based on the idiosyncratic level of attractiveness of the environment for criminal activity. Figure 1 shows an example of finding the width h of an h -spot from $J(x)$ and a given value for w .

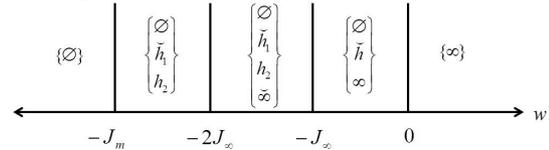
To obtain a taxonomy of equilibrium solutions for varying idiosyncratic attractiveness levels of the environment we note that there are three cases to con-

sider. In the following diagrams for each of the cases we show the sets of equilibrium solutions for various levels of idiosyncratic attractiveness w . Those equilibrium values which are not stable are indicated with a \checkmark symbol above the spot width. Stability of these solutions is shown in the next section.

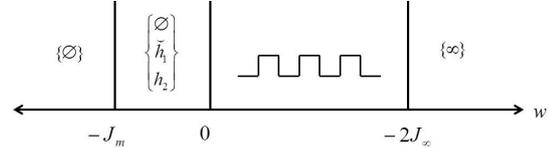
Case I_1 : $2J_{\infty} > J_m > 0$



Case I_2 : $J_m > 2J_{\infty} > 0$



Case 2: $J_{\infty} < 0$



We note for certain choices of w in all cases the field admits a bi-stability where both the quiet spot and hot spot solutions are stable. When $J_{\infty} < 0$ and $0 < w < -2J_{\infty}$ there is a multi-peak solution and no localized solution.

As noted in theorem 1, a quiet spot required $w < 0$. We can see from the taxonomy that if the idiosyncratic attractiveness is sufficiently negative a quiet spot is the only equilibrium solution. This would correspond to an environment where there is a high risk for committing a criminal activity. Achieving such an environment would be potentially resource intensive and likely too expensive to implement. As w is increased from these high risk values stable hot spot solutions become possible. Once the idiosyncratic attractiveness promotes crime, we find cases where the only equilibrium is one where crime persists everywhere in the environment. For stable hot spot solutions, it is required that the idiosyncratic attractiveness of the environment represent a risk for committing a criminal act. A risk that is enough to discourage crime without creating an environment where no crime exists may be accomplished through many guardianship mechanisms ranging from neighborhood watch programs to the criminal Justice system.

STABILITY

Hot spots of criminal activity are persistent in time implying that the h -spot solutions we have found should be stable equilibrium solutions. In this section we establish the stability of equilibrium solutions of (3) against perturbations of these equilibrium. The stability analysis performed by Amari developed a differential equation to described the evo-

lution of the width of an equilibrium solution. Then the stability of the width of an equilibrium solution is used as a proxy for the stability of the equilibrium solution of the original system. Here, we establish the stability of the equilibrium solution by a linearization of (3) around an equilibrium solution adapting the analysis of Blomquist *et al* (Blomquist et al., 2005) to a single state variable. Before engaging this analysis we note that for quiet spots $A(x) = w$ are considered stable. Additionally, solutions can only grow to ∞ -spots if $J_\infty + w > 0$; while such solutions can not be properly perturbed, *inf*ty-spots with $J_\infty < -w$ are not considered stable.

Let $A_e(x)$ denote an equilibrium solution of (3)

$$A_e(x) = \int_{-\infty}^{\infty} j(x-x') \Lambda(A_e(x')) dx'.$$

We consider a perturbed state of $A_e(x)$

$$A(x, t) = A_e(x) + \chi(x, t). \quad (8)$$

For the form of the acting out function given in (2) a Taylor expansion about the equilibrium solution yields

$$\Lambda(A_e + \chi) = \Lambda(A_e) + \delta(A_e) \chi + \dots \quad (9)$$

where $\delta(x)$ denotes a Dirac delta function and it is assumed $|\chi| \ll |A_e|$ so we may keep only the first two terms of the Taylor expansion. Plugging the perturbed state (8) into (3) and using the Taylor series approximation (9) we deduce a non-local evolution equation for the perturbation $\chi(x, t)$

$$\frac{\partial \chi}{\partial t} = -\chi(x, t) + \int_{-\infty}^{\infty} j(x-x') \delta(A_e(x')) \chi(x') dx' \quad (10)$$

We seek solutions of (10) of the form

$$\chi(x, t) = e^{\lambda t} \chi_1(x) \quad (11)$$

where the eigenvalues λ determine the the disturbance is growing ($\lambda > 0$) or decaying ($\lambda < 0$) indicating an unstable or stable equilibrium respectively. Plugging (11) into (10) we find

$$(1 + \lambda) \chi_1(x) = \int_{-\infty}^{\infty} \delta(A_e(x')) \chi_1(x') dx'.$$

Noting that for a h -spot $|\frac{d}{dx} A_e(0)| = |\frac{d}{dx} A_e(h)| = |j(0) - j(h)|$,

$$(1 + \lambda) \chi_1(x) = \frac{[j(x) \chi_1(0) - j(x-h) \chi_1(h)]}{|j(0) - j(h)|}. \quad (12)$$

We refer the reader to appendix B of Blomquist *et al* (Blomquist et al., 2005) for details of this computation. Evaluating (12) at the end points of the h -spot $x = 0$ and $x = h$, achieves the system of homogeneous linear equations, $M \vec{\chi}_1 = \vec{0}$

$$\begin{bmatrix} (1 + \lambda) - M_1 & M_2 \\ -M_2 & (1 + \lambda) + M_1 \end{bmatrix} \begin{bmatrix} \chi_1(0) \\ \chi_1(h) \end{bmatrix} = \vec{0}$$

where,

$$M_1 = \frac{j(0)}{|j(0) - j(h)|}, \quad M_2 = \frac{j(h)}{|j(0) - j(h)|}.$$

As we consider non-trivial perturbations $\chi_1(x)$, we find the λ by solving $\det(M) = 0$

$$\lambda_{+,-} = \frac{\pm \sqrt{j^2(0) - j^2(h)}}{|j(0) - j(h)|} - 1.$$

The root λ_- is always negative while the root λ_+ is negative when $j(h) < 0$ and positive when $j(h) > 0$. This means that h -spot solutions with $j(h) > 0$ exhibit a saddle point instability and those with $j(h) < 0$ are stable. Unstable solutions are noted in the equilibrium solution taxonomy tables by a \checkmark symbol over the corresponding spot width.

The stability of an h -spot solution depends on both the width h , and the social interaction function $j(x)$. For example we see from figure 1 that in cases where we find two h -spot equilibrium, the narrower h -spot corresponds to $j(h) > 0$ and is unstable while the broader h -spot corresponds to $j(h) < 0$ and is stable. The lack of stability of a narrow h -spot means that these spots can be dissipated by a small deterrence; whereas, dissipation of a broad h -spot would require a significant deterrence effort.

DISCUSSION

The routine activity theory makes micro and macro level assertions. On the micro-level the theory states that the convergence of potential victims and offenders in the absence of capable guardianship against the crime may lead to the emergence of crime. At a macro-level the theory asserts that features of the larger society and community or environment, may make these convergences more likely (Felson, 2008). The model presented here quantifies the capability of the guardian against the crime by an idiosyncratic attractiveness level w . In this case a lack of capable guardianship would be quantified be a positive idiosyncratic attractiveness that could potentially lead to rampant crime equilibrium solution consistent with the routine activity theory. The presence of a guardian is important to deter the occurrence of crime events; however, the presence of a guardian may not be sufficient to ensure that there is no crime. The level of capability of the guardian to deter the crime is quantified by negative idiosyncratic attractiveness. The presence of the hot-spot solutions that emerge from our model require the presence of a level of guardianship that can deter rampant crime; yet, is not so strong a deterrent as to prohibit all crime. Guardianship against crime acts can take many forms ranging from policing to alarm systems to the implementation of the criminal justice system for example. The ubiquity of hot-spots of criminal activity would imply that currently implemented levels of such guardianship are in

this range of deterring rampant crime solutions while not quieting all crime. A dissipation of the hot spot would require a temporary increase in the guardianship to decrease the width of the hot-spot to below the narrow width or unstable width hot-spot solution for the idiosyncratic background level. However, to protect against the resurgence of hot-spots of criminal activity would require maintaining a high level of guardianship. Such high levels of guardianship may be cost prohibitive to implement and require levels of guardianship that are socially unacceptable. While the success of the implementation of a surge and maintain strategy to dissipate and protect against the re-emergence of crime hot-spots remains to be seen in practice, Brazil appears to be implementing such a strategy in its preparation for the 2014 World Cup (Associated Press, 2014).

In this paper we have presented a model for the non-local aggregation of environmental attractiveness for criminal activity via a social interaction mechanism. This social interaction takes the form of an opinion dynamic mimicking a voter model. In a voter model a binary decision or vote is made by one agent and neighboring agents may be influenced to change their vote based on this vote (Xia et al., 2011). In this case the vote is for the occurrence of crime based on local attractiveness that is influenced by non-local votes for crime. The influence to vote for or against crime in this manner is distance dependent via a social interaction weighting function.

We build on earlier work considering a social interaction model (Berestycki and Nadal, 2010) for the communication of risk in influencing the propensity of the offender to commit a crime activity. Rather than considering local costs for repeat victimization for a particular location, decreases in the attractiveness of the environment or deterrence for criminal activity at a location are communicated through the social interaction function. The decrease in attractiveness could be a consequence of, for example, the presence of guardians owing to the crime level or an attraction of offenders to a more attractive location for criminal activity. We analyze the existence and stability of hot spots of criminal activity as equilibrium solutions of the model when deterrent forces are non-local and attractive forces are non-constant across the whole domain. We find that hot spots exist so long as there is an overall risk in the environment for committing a criminal activity and that a bi-stability can occur between quiet and hot spot environments.

The model studied is a reformulation of Amari's seminal model for studying lateral inhibition type neural fields (Amari, 1977). We followed Amari's analysis to give necessary and sufficient conditions for the existence of different types of equilibrium solutions including hot spot solutions and provided a taxonomy of equilibrium solutions such as provided by Amari for neural fields. However, we departed

from Amari's approach in consideration of the stability of the equilibrium solutions. Our approach of a direct analysis of perturbations of equilibrium solutions for (3) provides more information about the nature of the instability that arises than Amari's approach of considering the stability of the spot width itself.

In this model we considered an environment where the inherent attractiveness of the environment is homogeneous $W(x, t) = w$. Amari considered for neural fields a stationary input stimulus, $W(x, t) = W(x) > 0$, and found that the localized activity states would be centered at local maxima of the input stimulus. In this context we could think of $W(x)$ as a function that has negative minima at areas of strong guardianship and potentially positive at areas where the environmental cues suggest a low risk for criminal activity such as in the broken windows theory (Keizer et al., 2008). We would expect in following the Amari analysis that hot spots would form around a positive local maxima of this stationary inherent attractiveness.

Additionally we have only considered the properties of one hot spot. Typically crime maps show a patterning of distinct separated crime hot spots. Amari considered the interaction of localized excitation patterns in neural fields. Following Amari's analysis we would expect in this model that for two crime hot spots that are sufficiently close the two crime hot spots will attract each other to form one crime hot spot. At a more intermediate distance the two hot spots will repel one another until they are sufficiently separated to have no influence on each other. The exact distance that hot spots separate will depend on the social interaction function.

A common policing strategy known as "cops on the dots" is to send police to the crime hot spots to provide a strong deterrence. The result of this strategy may be to either dissipate crime or locate it to another location (Braga, 2001). Recently using this strategy to dissipate a hot spot in a reaction diffusion model (Zipkin et al., 2013) and existence for traveling wave solutions in a reaction diffusion model for criminal propensity to act (Berestycki et al., 2013). Amari introduced a "two layer" version of his neural field model that mimics the cops on the dots strategy. In this extension the deterrent force would be considered a separate field variable that is excited only at the location of the crime and provides a spread of deterrent to the attractiveness field. Amari exhibited that such a two layer network can exhibit traveling wave solutions. The social interaction function presented here can be derived from such a two layer network when the deterrent layer is considered to be in a quasi-steady state with the description of the level of attractiveness.

An additional feature of crime hot spots for future consideration in this modeling framework is the influence of the topology of the environment on the for-

mation of crime hot spots. If an offender is unaware of a target location then the offender can not commit a crime at that location. The topology of crime concentration reflects a communication of potential targets and risks (Brantingham and Brantingham, 2008; Johnson, 2010) that may be reflected in the social interaction function. Explorations of neural field equations like that of Amari that incorporate the topology of the neural network have shown new non-trivial equilibrium and traveling wave solutions (Haskell and Bressloff, 2003; Haskell and Paksoy, 2011; Salomon and Haskell, 2012, 2013). It would be interesting to further study these field equations integrating a topology that arises from environmental criminology for continued comparison and advancement of the understanding of the influence of network topology on crime hot spots.

REFERENCES

- Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, 27(2):77–87.
- Associated Press (2014). Brazilian security forces prepare to ‘pacify’ new rio slum district. <http://www.theguardian.com/world/2014/mar/30/rio-favela-security-pacify-world-cup>.
- Berestycki, H. and Nadal, J.-P. (2010). Self-organised critical hot spots of criminal activity. *European Journal of Applied Mathematics*, 21(4-5):371–399.
- Berestycki, H., Rodriguez, N., and Ryzhik, L. (2013). Traveling wave solutions in a reaction-diffusion model for criminal activity. *Multiscale Modeling & Simulation*, 11(4):1097–1126.
- Blomquist, P., Wyller, J., and Einevoll, G. T. (2005). Localized activity patterns in two-population neuronal networks. *Physica D: Nonlinear Phenomena*, 206(3):180–212.
- Braga, A. A. (2001). The effects of hot spots policing on crime. *The ANNALS of the American Academy of Political and Social Science*, 578(1):104–125.
- Brantingham, P. and Brantingham, P. (2008). Crime pattern theory. In Wortley, R. and Mazerolle, L., editors, *Environmental Criminology and Crime Analysis*. Routledge.
- Brantingham, P. J. and Brantingham, P. L. (1981). *Environmental criminology*. Sage Publications Beverly Hills, CA.
- Chaturapruek, S., Breslau, J., Yazdi, D., Kolokolnikov, T., and McCalla, S. G. (2013). Crime modeling with lévy flights. *SIAM Journal on Applied Mathematics*, 73(4):1703–1720.
- Cohen, L. E. and Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608.
- Cornish, D. B. and Clarke, R. V. (2008). The rational choice perspective. In Wortley, R. and Mazerolle, L., editors, *Environmental Criminology and Crime Analysis*. Routledge.
- Felson, M. (2008). Routine activity approach. In Wortley, R. and Mazerolle, L., editors, *Environmental Criminology and Crime Analysis*. Routledge.
- Haskell, E. C. and Bressloff, P. C. (2003). On the formation of persistent states in neuronal network models of feature selectivity. *Journal of integrative neuroscience*, 2(01):103–123.
- Haskell, E. C. and Paksoy, V. E. (2011). Localized activity states for neuronal field equations of feature selectivity in a stimulus space with toroidal topology. In *Nonlinear and Complex Dynamics*, pages 207–216. Springer New York.
- Johnson, S. D. (2010). A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics*, 21(4-5):349–370.
- Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., and Townsley, M. (2007). Space-time patterns of risk: a cross national assessment of residential burglary victimization. *Journal of Quantitative Criminology*, 23(3):201–219.
- Keizer, K., Lindenberg, S., and Steg, L. (2008). The spreading of disorder. *Science*, 322(5908):1681–1685.
- Salomon, F. and Haskell, E. C. (2012). Travelling wave solutions for ring topology neural fields. In Vigo-Aguiar, J., editor, *12th International Conference Computational and Mathematical Methods in Science and Engineering*, pages 1523–1531.
- Salomon, F. and Haskell, E. C. (2013). On spatio-temporal patterns in two-layer ring topology neural fields. In Rekdalsbakken, W., Bye, R. T., and Zhang, H., editors, *ECMS*, pages 870–876. European Council for Modeling and Simulation.
- Short, M. B., Brantingham, P. J., Bertozzi, A. L., and Tita, G. E. (2010). Dissipation and displacement of hotspots in reaction-diffusion models of crime. *Proceedings of the National Academy of Sciences*, 107(9):3961–3965.
- Short, M. B., D’Orsogna, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A. L., and Chayes, L. B. (2008). A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267.
- Weisburd, D., Bruinsma, G. J., and Bernasco, W. (2009). Units of analysis in geographic criminology: historical development, critical issues, and open questions. In *Putting crime in its place*, pages 3–31. Springer.
- Xia, H., Wang, H., and Xuan, X. (2011). Opinion dynamics: A multidisciplinary review and perspective on future research. *International Journal of Knowledge and Systems Science*, 2(4):72–91.
- Zipkin, J. R., Short, M. B., and Bertozzi, A. L. (2013). Cops on the dots in a mathematical model of urban crime and police response.

AUTHOR BIOGRAPHIES

EVAN HASKELL is an Associate Professor of Mathematics at Nova Southeastern University. Dr. Haskell holds a Ph.D. in Mathematics from the Courant Institute of Mathematical Sciences, New York University. Previously Dr. Haskell was a visiting member of Cognitive Neuroscience sector at the International School for Advanced Study (SISSA) in Trieste, Italy, Scott Assistant Professor of Mathematics at University of Utah, and Visiting Assistant Professor of Mathematics at College of William and Mary. Dr. Haskell has published numerous articles in computational neuroscience.

THE DEFINITION OF STRESS SITUATIONS AND THEIR PREDICTION USING LIQUIDITY IN THE FRAMEWORK OF THE EMIR REGULATION

Barbara Dömötör
Kata Váradi, Ph.D.
Department of Finance
Corvinus University of Budapest
1093, Budapest, Hungary
E-mail: barbara.domotor@uni-corvinus.hu

KEYWORDS

EMIR regulation, Value at Risk models, market liquidity measurement, stress definition

ABSTRACT

The role of the central counterparties (CCP) in the financial sector is very important, since they bear the counterparty risk during the trading on stock exchanges. Because of the notable risk central counterparties have to face, the attention of the regulators has turned towards them lately, by defining several processes how the CCPs should measure and manage their risk. The definition of stress has a crucial role, however it is not specified clearly. Based on the regulation, we investigate a possible definition of stress, its consequences on the Hungarian stock market, and its relationship to and predictability from market liquidity.

INTRODUCTION

Trading with financial instruments takes place on stock exchanges or on over-the-counter (OTC) markets. One of the basic differences between the operation of the two markets is the presence of the central counterparty on the organised markets, that acts as the trading partner in each trade. The role of the CCP is to take over counterparty risk, namely the risk that one of the parties will not perform as promised (Brealey et al. 2011). Since the counterparty risk is a notable risk category, CCPs should measure and manage their risk efficiently, in order to maintain the market's financial stability, as bankruptcy of a CCP would have a serious effect on the whole financial sector. It is also important to note, that in the future, regulators aim to extend the activity of the CCPs for the OTC markets as well, in order to decrease the risks on that market segment also.

As a regulatory answer for the financial crises, on 4th July, 2012 the European Parliament and the Council established a new regulation, called European Market Infrastructure Regulation (EMIR, Regulation (EU) No. 648/2012 on OTC derivatives, central counterparties and trade repositories). This regulation was supplemented by the European Commission on 19th December, 2012, with the Regulation (EU) No

153/2013, providing the technical standards of the EMIR regulation.

The above regulations aim to ensure the prudence of the risk management procedure of the central counterparties, however in some cases the EMIR is not specific enough, and doesn't give an exact solution how to interpret some notions, like one of its key terms: the stress situation. The proper definition is important, since according to the regulation, the different applications of certain models are based on whether a financial stress is present on the market or not.

Although the role of the central counterparties and their regulation has an increasing literature recently, the problem of defining stress has arisen as a practical issue yet, so we do not know any academic study dealing with it.

In this paper we present a definition for the stress situation, based on the results of the backtest of the applied risk measurement model, and we show its effect on real market data of the Hungarian Stock Exchange, in the after crisis period, between 2010 and 2013. Furthermore, we analyse how the identified stress situation would have been predictable from the (i)liquidity of the market.

Our paper is built up as following: first, we introduce the regulations focusing on their elements that apply the notion of stress. Then we introduce the risk measure Value-at-Risk (VaR), as the risk management process and our analysis is based on it. The next section presents the measurement of market liquidity, since the liquidity of the market in stress is tested also. The following part contains our empirical research, the methodology, the market data and the analysis of the stress situation and its co-movement with liquidity. The last section summaries our results and conclusions.

THE REGULATIONS

The main risk, CCPs are facing, derives from the default of their clearing parties. For taking over this risk, CCPs apply a waterfall system of collateral elements, that decreases their losses if one of the parties does not fulfil its obligations. The first component ensuring the performance of the trading parties is an initial margin, a certain amount of cash or cash-equivalent that is required to be placed by both

parties - the seller and the buyer - of the trade. The concept of determining the level of this margin is regulated by EMIR and the 153/2013 regulation.

Model of margin determination

The regulation says, that 'a CCP shall calculate the initial margins to cover the exposures arising from market movements for each financial instrument that is collateralised on a product basis' (Regulation 153/2013/EU, Article 24). The regulators do not define the models the CCPs shall use. The only limitations they give are the following:

1. CCP has to use a 99% confidence interval in case of financial instruments other than OTC derivatives, and 99.5% for OTC derivatives (Regulation 153/2013/EU, Article 24).
2. For estimating the model CCPs shall use the data at least of the latest 12 months' data (Regulation 153/2013/EU, Article 25).
3. CCPs shall take into account the time horizon for the liquidation period, which shall be two days for financial instruments other than OTC derivatives, and five days for OTC derivatives (Regulation 153/2013/EU, Article 26).

The most widespread risk measure used also by the Basel rules – regulating financial institutions – is the Value-at-Risk (VaR), and because of its popularity and applicability for the purposes of the EMIR, it was applied by most of the CCPs, too. The features, shortcomings and alternatives of VaR are introduced in details in the next section.

Time horizon for historical volatility

The regulation requires a period of at least 12 months to be used to estimate the historical volatility. Besides that the regulation requires further specification in order to be prepared for even extreme market circumstances, by prescribing a 'full range of market conditions, including periods of stress' (Regulation 153/2013/EU, Article 25).

This means, that the definition of stress has an effect on the observation period the CCP shall use to calculate the model, and also on the calculated volatility and margin level.

Procyclicality

The financial crisis shed light on the possible procyclical effect of the risk management regulations in the financial sector. The models using more rigorous capital requirements in case of market turbulences, contributed to the financial difficulties of the institutions and deepened even the crisis. Consequently the latest direction of the macroprudential regulations aims to minimise that effect by applying anticyclical provisions.

That was formulated in Regulation 153/2013/EU: 'A CCP shall ensure that its policy for selecting and revising the confidence interval, the liquidation period and the lookback period deliver forward looking, stable and prudent margin requirements that limit procyclicality to the extent that the soundness and financial security of the CCP is not negatively affected' (Regulation 153/2013/EU, Article 28). To achieve this goal, the regulator requires the CCPs to use a margin buffer at least equal to 25%, when calculating margin in normal market conditions. On the other hand, in case of changing market conditions, which would cause an essential rise in the margin requirements, the CCP can disregard the margin buffer. This procedure is to be used in stress situation that is depending on the definition of stress, also.

Definition of stress

The present practice of most CCPs relies on the decision of the risk management committee when deciding about the existence of stress. Although we agree to maintain this kind of flexibility, it is suggested to define some objective criteria that give a signal that the market may be regarded in stress.

As the risk measurement models are reviewed and tested on a daily basis, we suggest to use the results of these backtests as a warning signal about stress¹. If the real market change exceeds the maximal movement based on the applied risk measure (VaR) for one or more main products, the situation is to be analysed further as stress situation is assumable.

RISK MEASURING MODELS

A risk measure is defined as a function that assigns a scalar to a random variable quantifying a certain loss. In the models of the capital market, standard deviation is used to quantify risk, but for risk management purposes measures focusing on the downside outcomes are more appropriate.

Value-at-Risk (VaR) was defined by JP Morgan in the mid 90's, as the maximum loss of the portfolio over a predefined time horizon (T) at a given significance level (α) under normal circumstances. VaR can be expressed either in absolute value or as a percentage of the portfolio value (Jorion, 2007). Because of its simplicity VaR was adopted by the whole financial sector, even though the regulation of the financial institutions – Basel Rules – uses it in the risk management systems, since it is easy to use and to understand.

In order to calculate VaR, the probability distribution of the position in the certain security/portfolio at time T is to be modelled.

¹ We thank for the idea Edina Berlinger, who lead the risk management validation project of the Hungarian CCP.

The $(1-\alpha)$ th percentile of this distribution shows the threshold value (K) the portfolio underperforms with a probability of $(1-\alpha)$ at time T (Jorion, 2007), as it is shown in Equation (1).

$$P(V_t < K) = 1 - \alpha \quad (1)$$

where the value of the position is V at time t .

Value-at-Risk is given as the difference of the actual value and the threshold.

There are 3 main concepts to calculate VaR: historical calculation, analytical method and Structured Monte Carlo Simulation (Jorion, 2007). In the framework of the historical method, the events of a chosen reference period are supposed to describe the potential future outcomes, so the whole distribution is given by them. The analytical method assumes the knowledge of the distribution, and as it is provided in most of the cases to be normal, this method is also called delta-normal method.

The third possibility to determine the distribution is simulation that can rely either on historical data or on the knowledge of the value generating process.

Although VaR is not a coherent risk measure, as it was presented by Artzner et al. (1999), and a coherent alternative was suggested by Acerbi and Tasche (2002), it is still the most commonly used risk measure in the financial sector. Even if EMIR does not restrict the circle of the applied risk measures to VaR, most CCPs use VaR for risk management, to measure their risk and to calculate margin requirements.

MEASUREMENT OF LIQUIDITY

The predictability of financial difficulties or even crisis would be very important for both micro- and macroeconomic perspectives. As financial stress is often attended by liquidity shortages, the question arises, whether liquidity can be used as an indicator of the forthcoming stress.

The notion of liquidity has several interpretations, like the liquidity of a company, the liquidity of the whole financial system, or the liquidity of the market. In each different interpretation liquidity is to be measured differently and so the management of illiquidity risk differs too, that is why it is always very important to clarify which liquidity notion we are using. In our paper we focus on the concept of market liquidity. The Bank for International Settlements (BIS) gave a generally accepted definition for market liquidity: 'Liquid markets are defined as markets where participants can rapidly execute large-volume transactions with little impact on prices.' (BIS, 1999)

The definition of liquidity suggests that its concept is very complex. There does not even exist a single best

way to measure its value. A broad overview of liquidity indicators is provided by von Wyss (2004). The liquidity indicators can be grouped into three main categories (Csavas and Erhart, 2005): (1) indicators of transaction costs, (2) indicators of volumes, (3) indicators of prices.

In our analysis we will focus on a liquidity indicator, that is based on transactions cost, the so called Budapest Liquidity Measure (BLM). Since the notion of BLM is quite new in the literature of Finance, we introduce it in the next sub-section.

Budapest Liquidity Measure

The Budapest Liquidity Measure (BLM) belongs to the class of liquidity measures. The first liquidity measure of this type was the Xetra Liquidity Measure (XLM) created by the Deutsche Borse Group in 2002, by Gomber and Schweikert (2002). The same measure was introduced on the Budapest Stock Exchange (BSE) under the name of Budapest Liquidity Measure (BLM) in 2005 (Kutas and Vegh, 2005, Gyarmati et al. 2010). These liquidity measures are weighted spread measures that represent the implicit costs of trading, which arise from the fact that actual trading is not executed at the mid-price. The simpler version of the liquidity measures is the relative spread measure, which can be computed according to Equation (2):

$$RS_{spread}_t = \frac{p_t^{Ask} - p_t^{Bid}}{\frac{1}{2}(p_t^{Ask} + p_t^{Bid})}, \quad (2)$$

where p_t^{Ask} denotes the best ask, and p_t^{Bid} the best bid price in the order book at time t . This measure displays the loss realized when buying and then immediately selling the same asset, relative to the mid-price (average of the best bid and ask price in the order book).

Basically, the BLM is a version of the relative spread measure. The difference is, that in case of the relative spread, only the best ask and bid price appear in the calculation, while in case of BLM we take into account that an order is not necessarily fulfilled on the best price levels. The calculation of BLM is shown in Equation (3):

$$BLM_t = \frac{\sum_i p_{t,i}^{Ask} \cdot q_{t,i}^{Ask} - \sum_i p_{t,i}^{Bid} \cdot q_{t,i}^{Bid}}{p_t^{Mid} \cdot q_t} \quad (3)$$

where $p_{t,i}^{Ask(Bid)}$ shows the i th best price on the ask(bid) side at time t , whereas $q_{t,i}^{Ask(Bid)}$ denotes the depth of (the overall quantity submitted to) that same price level. p_t^{Mid} is the mid-price at time t .

An interpretation of the calculation of BLM is shown on Figure (1).

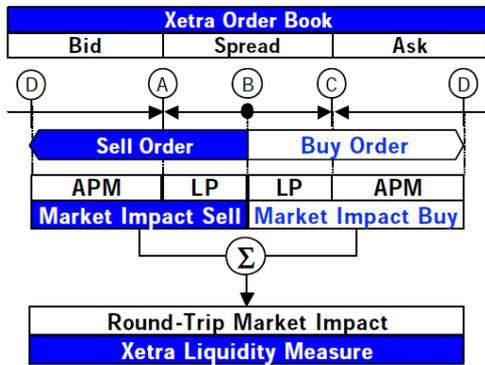


Figure 1: The Calculation of the Liquidity Measure
 Source: Gomber and Schweikert (2002), p. 3.

In sum, the smaller these measures, the higher the liquidity of the asset.

The Budapest Stock Exchange calculates this measure every time when there is a change in the order book, and also on a daily basis. The daily BLM value is the average of the intraday BLM data. The BSE publishes these daily data at the end of every month in order to provide information for the market.

EMPIRICAL TEST OF STRESS

In this research, first we test our stress-definition on real market data of the Budapest Stock Exchange. We calculate risk measure (VaR) for every day and investigate the potential stress signals in the analysed period.

For the analysis we used the daily closing prices of the bluechip stocks of the Budapest Stock Exchange, namely, the OTP, MOL, Richter, MTelekom of the last 4 years, between January 2010 and December 2013. In Finance the daily (log)return of financial assets are regarded to be a stationer random variable whose realisations derive from independent, identical distribution. Despite of some stylised facts (e.g. fat tail phenomena), daily logreturn is assumed to be normally distributed in most of the models. Following the literature, we calculate the daily logreturn (y) of the stocks, according to the Equation (4):

$$y_t = \ln\left(\frac{S_t}{S_{t-1}}\right) \quad (4)$$

where S denotes the stock price and the indices stand for the time.

The Value-at-Risk for each day is calculated according to the delta-normal method, the parameters of the return generating process are calculated as the average

(μ) and standard deviation (σ) of the logreturns in the previous 250 days, as it is shown in Equation (5):

$$VaR = \mu + \Phi^{-1}(1 - \alpha) * \sigma \quad (5)$$

where Φ^{-1} denotes the inverse of the cumulated distribution function of the standard normal distribution.

After having the logreturn and the VaR for every day, those days are to be investigated further, where the real price fall exceeds the VaR of the previous day, that should happen in $1 - \alpha$ percent of the cases, as a consequence of the VaR-definition. We used a significance level (α) of 99%, as it is prescribed by EMIR. Figure (2) illustrates the calculation in case of MOL, the Hungarian Oil company, representing almost one third of the Hungarian stock market.

The points below the red line show the days, when the negative price movement exceeded the maximal loss predicted by VaR with a probability of 99%. These days are to be examined further in order to decide about stress, according to our suggestion.

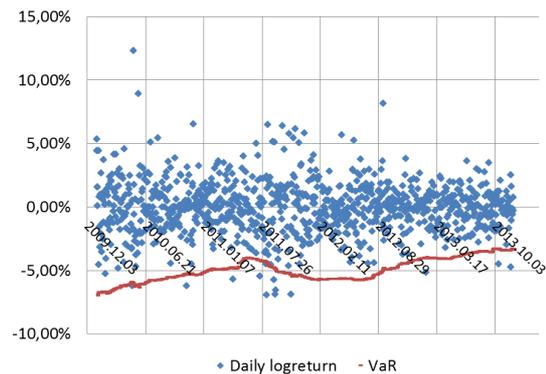


Figure 2: The daily logreturns and VaR of MOL between 2010 and 2013.

Source: own calculation based on the data of BSE.

As the outlying points can be caused by company specific reasons, we searched for the outlying days for all 4 stocks, in order to find those periods, when more assets give a warning signal. According to the VaR model the number of the outlying days should sum up to 1 percent, so 10 days out of the 1000 working days during the period. We found 8-15 outliers for each of the tested group of stocks – the least, 8 in case of MTelekom, and the most, 15 in case of OTP. This result supports the applicability of the VaR model, as even in case of MOL, the difference is insignificant.

The potential stress days are depicted on Figure (3).

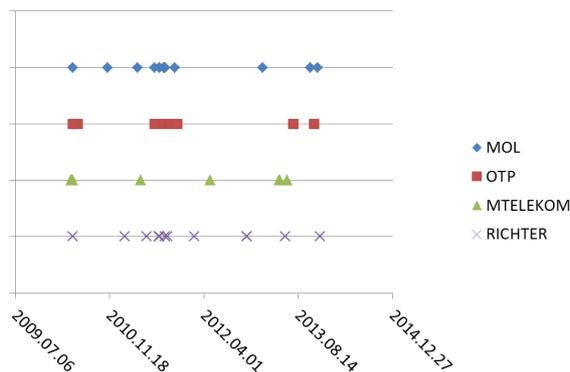


Figure 3: Stress signals of 4 Bluechips between 2010 and 2013.

Source: own calculation based on the data of BSE.

The only date when all the 4 tested papers signalled was the 6th of May 2010, when a sudden fall on the New York Stock Exchange due to technical problems caused worldwide market turbulences. Even then the market calmed down rapidly, so it is not reasonable to speak about a stress period.

Two other dates are worth investigating in the period, because 3 of the 4 stocks (except for MTelekom) signalled. The market fall of 8-9th of August 2011 was a consequence of global markets' events. In September 2011 the possibility to pay back foreign denominated mortgage loan at out-of market exchange rate forced to the Hungarian banking sector caused the extreme price movement, but these days were also followed by some correction, that offset the losses of the previous market collapse.

The further warning dates are triggered by single stocks, so they are not to be interpreted as stress period in the market.

Based on the above, we can state, that the examined period of the last 4 years was characterised by quiet market movements and free of stress.

THE TEST OF LIQUIDITY IN STRESS

Even if we have not found evidence of a real stress, it is worth to analyse the market liquidity in the periods of the signalling days.

For the purpose to quantify liquidity we used the daily value of the Budapest Liquidity Measure, for the same stocks and for the same time period as in the case of the daily logreturn calculation. The time series are given by the Budapest Stock Exchange. BLM refers to the cost of trading a certain amount, expressed in basis points, in the calculations we used the 20.000 euro BLM figures, referring to the cost of trading in that volume. The time series of BLM need to be differentiated also in order to get stationer data, consequently we calculated the daily change of BLM.

As the liquidity shortage is indicated by growing BLM figures, similarly to the stress calculation, we looked for those days, when the daily change exceeded the 99% maximum of the previous 1 year period. We had access to BLM figures from 2010, so the analysed time-series shortened to 3 years because of the reference period.

For the purpose to quantify liquidity we used the daily value of the Budapest Liquidity Measure, for the same stocks and for the same time period as in the case of the daily logreturn calculation. The time series derive from the Budapest Stock Exchange.

The delta normal method cannot be applied, since the daily differences of BLM are not to be regarded normal, as shown on the example of MOL on Figure (4). The rejection of normality was confirmed by the Kolmogorov-Smirnov test also (with a p -value of 0,000).

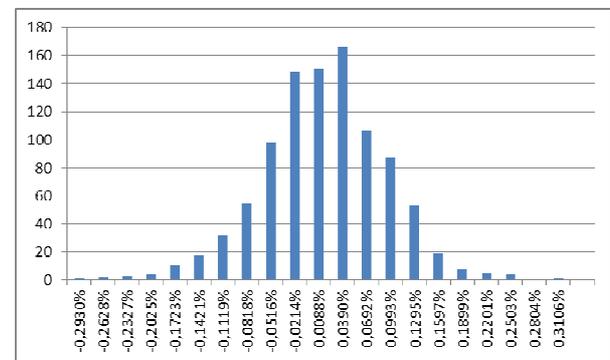


Figure 4: Distribution of the daily BLM differences of MOL between 2010-2013.

Source: own calculation based on the data of BSE.

Therefore, we applied the historical method to calculate Value-at-Risk characteristic risk measure for BLM, too. We took the 99% percentile of the data, and warning signal was defined as those days, when the daily change of BLM exceeded the 99% percentile of the previous 250 days.

First, we examined the signals of the liquidity measure in the two periods – August and September 2011 - identified by stock returns.

OTP		MOL	
Stock_return	BLM	Stock_return	BLM
22.09.2011		09.09.2011	
21.09.2011		08.09.2011	
20.09.2011	12.09.2011	02.09.2011	
08.09.2011	09.09.2011		
17.08.2011	18.08.2011	09.08.2011	11.08.2011
09.08.2011		08.08.2011	09.08.2011

Table 1: Signalling dates of MOL and RICHTER between 2011-2013.

Source: own calculation based on the data of BSE.

The BLM of RICHTER gave no signals at all, and MTELEKOM had no extreme price fall in the period, so only the stress dates of MOL and OTP are shown in Table 1. We can see, that merely about the half of the warning dates were accompanied by a liquidity signal, and even in these cases the liquidity measure signs followed the market fall, instead of predicting it. It seems as if market participants withdraw their orders after the price fall of the market and not the reduction of the order book causes the fall of the prices.

The other warning signals of BLM in the period appeared independently from the extreme market movements.

The reason of our results could be, that there were no real stress in the last 4 years that explains the independence of the highest changes in price and liquidity.

CONCLUSION

Based on the recent direction of the regulation of financial markets and institutions, in this paper an objective reference for defining stress situation was suggested. As our empirical analysis showed, this basis has also some subjective elements, and further investigation of the market is needed in order to decide whether stress exists. We have found 3 dates in the reference period, when at least 3 of the 4 analysed stocks alarmed for stress simultaneously.

The paper analysed furthermore the connection between the above defined stress signal and market liquidity. We have found no strict connection of the price and liquidity movement. In contrast to our expectations the liquidity shortage rather followed the extreme price changes, than predicted it.

The market movements of the tested period – between 2010 and 2013 – proved to be very quiet that can explain our results. The analysis is to be extended for a longer period containing the years of the financial crisis as well.

REFERENCES

- Acerbi, C., Tasche, D. (2002): Expected Shortfall: A Natural Coherent Alternative to Value at Risk. *Economic Notes* 31 No. 2. July, pp. 379–388.
- Artzner, P., Delbaen, F., Eber, J. M., Heath, D. (1999): Coherent Measures of Risk. *Mathematical Finance* Vol. 9, No. 3. pp. 203–228.
- Bank for International Settlements (1999): Market Liquidity: Research Findings and Selected Policy Implications. *Committee on the Global Financial System, Publications*, No. 11.
- Brealey, R.A., Myers, S.C., Allen, F. (2011): *Principles of Corporate Finance*. McGraw-Hill Companies, Inc. 10th Edition.
- Csávás, Cs., Erhart, Sz. (2005): Likvidek-e a magyar pénzügyi piacok? – A deviza- és állampapír-piaci likviditás elméletben és gyakorlatban (Are the Hungarian Money Markets Liquid? – The Liquidity of the Foreign Currency- and the Government Security Market in

Theory and in Practice). *Hungarian National Bank working paper* No. 44.

- Gomber, P., Shcweikert, U. (2002): The Market Impact – Liquidity Measure in Electronic Securities Trading. *Die Bank*, 7/2002.
- Gyarmati, Á., Michaletzky, M., Váradi, K. (2010): Liquidity on the Budapest Stock Exchange 20, 07-2010. Budapest Stock Exchange, working paper. Available at: <http://ssrn.com/abstract=1784324>
- Jorion, P. (2007): *Value at Risk*. McGraw-Hill Companies, Inc. 3rd Edition.
- Kutas, G., Végh, R. (2005): A Budapesti Likviditási Mérték bevezetéséről (About the Introduction of the Budapest Liquidity Measure). *Economic Review*, Volume LII, July-August, pp. 686-711.
- Von Wyss, R. (2004): *Measuring and predicting liquidity in the stock market*. Universität St. Gallen, Dissertation.
- Regulations:
- European Market Infrastructure Regulation: Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4 July 2012, on OTC derivatives, central counterparties and trade repositories. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:201:0001:0059:EN:PDF>
- Supplementation of EMIR: Commission Delegated Regulation (EU) No 153/2013 of 19 December 2012, supplementing Regulation (EU) No 648/2012 of the European Parliament and of the Council with regard to regulatory technical standards on requirements for central counterparties. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:052:0041:0074:EN:PDF>

AUTHOR BIOGRAPHIES

BARBARA DÖMÖTÖR is an Assistant Professor of the Department of Finance at Corvinus University of Budapest. Before starting her PhD studies in 2008, she worked for several multinational banks. She is now working on her doctoral thesis about corporate hedging. She is lecturing Corporate Finance, Financial Risk Management and Investment Analysis, her main research areas are financial markets, financial risk management and corporate hedging. Her e-mail address is: barbara.domotor@uni-corvinus.hu

KATA VÁRADI is currently an assistant professor at the Corvinus University of Budapest (CUB), at the Department of Finance. She graduated also at the CUB in 2009, and after it obtained a PhD in 2012. Her main research area is market liquidity – the topic of her PhD dissertation was the liquidity of the stock markets –, but she also does researches related to the bonds markets and capital structure of companies. Her e-mail address is: kata.varadi@uni-corvinus.hu

ACKNOWLEDGEMENTS

This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP-4.2.4.A/ 2-11/1-2012-0001 ‘National Excellence Program’ for Kata Váradi, PhD.

PATH DEPENDENCY IN INVESTMENT STRATEGIES – A SIMULATION BASED ILLUSTRATION

Ágnes Vidovics-Dancs
Péter Juhász, PhD, CFA
János Száz, CSc

Department of Finance
Corvinus University of Budapest
H-1093, Fővám tér 8, Budapest, Hungary
E-mail: agnes.dancs@uni-corvinus.hu

KEYWORDS

Path dependency, leverage, financial simulation, risk

ABSTRACT

In finance, the term path dependency is typically used when valuing derivative assets like American or Asian type options. Our simulation based example illustrates that the final payoff of an investment strategy could also depend on the previous historical price movements of the asset in our portfolio even if the final selling price of the asset itself is independent of it.

As illustration, we use the real life monthly return data of the shares of the Hungarian oil company (MOL), and we show that it does matter what path the stock price follows from the purchase to the date of selling if we finance our portfolio from a debt requiring regular payments throughout the holding period. In our model the investor covers the required cash outflows by selling some of the shares originally bought. Over a ten year period one may achieve a total return between -100.0 and 1,026.0 per cent depending on the path of the share quotation generated randomly by mixing real life monthly returns. In 7.95 per cent of the cases we would even go bankrupt before the 10 years are over.

INTRODUCTION

In the financial literature the term path dependency is used in different contexts. Sometimes, just as in the social sciences, we refer with this expression to the fact that our current or future payoffs, decisions or strategic options are limited or determined by our previous choices or our history (e.g. Graves (2011) on entrepreneurial finance in Southern states of US or Bianco et al. (1997) on financial systems). Pierson (2000) offers a conceptualisation of “path dependence” in this relation, while Dobusch and Kapeller (2013) contrast a number of recent articles showing the different approaches when using the term in this meaning.

In other cases we use a more narrow meaning and consider something path dependent, once the value of the given asset depends not only on the price of another

item at a set point in time, but also on the price dynamics of the other asset during a period of time. (E.g. Thompson (1995) on contingent claims, Baule and Tallau (2011) on bonus certificates or Jazaerli and Saporito (2013) on the path dependence of the Greeks of options.) In this article we use the expression in its later meaning, where it is not our historical decisions but a set of past events out of our control that influence our final payoff. This kind of definition in relation to investment in a project is very well presented and contrasted to real options by Adner and Levinthal (2004). The basic idea of this paper first appeared in the article of Száz (2013).

We examine a very stylized company with $E=1$ equity and D amount of debt. Taxes, dividends and transaction costs are ignored. Furthermore, our company does not trade, produce or offer services, but its activity is limited to investing in and holding of one given financial asset called *stock*. The only decision the company may and has to take is whether to buy the stock only using equity (*Strategy A*) or creating a leveraged portfolio (*Strategy B*). Next we describe the two strategies in detail.

Strategy A: This is our benchmark strategy. The firm takes no debt ($D=0$), and spends the total of its shareholder capital (the equity, E) on buying stocks. After T years we close the position and sell the stocks.

Strategy B: Taking some debt (D) with maturity of T years the company spends at the start the total of its own and borrowed capital ($E+D=V$) on buying stocks. Interest and principal payments of the debt have to be covered from selling the appropriate quantity of the stocks owned. After T years we close the position and sell the remaining stocks. Payments on the debt are made monthly, the interest rate is r per annum, and principal payments are made monthly in equal sums across the whole lifetime of the debt. For this article for the sake of simplicity we assume $D=2$.

The relative performance of the two strategies is obviously dependent on three factors:

- the price dynamics of the stock;
- the amount of debt (D) (level of leverage: $L=V/E$);
- the interest rate of the debt (r).

A REAL LIFE EXAMPLE

For comparing the two strategies, we have chosen the actual price dynamics of an existing stock, namely MOL (Hungarian oil and gas company). The period examined is 1998-2008, hence $T=10$ years. Figure 1 illustrates the price dynamics of MOL shares during this period. For easier overview we rescaled the data by choosing the initial stock price on the first day to be the basis (100%).

As one may observe in the first half of the period MOL was travelling horizontally without any relevant up- or downtrend. After the first half of 2003 a strong uphill began, but some high drops also occurred. In the 10-year period, the price of MOL shares multiplied by 4.5 so the average growth rate was 16.3% p.a. This means that investing in this stock would have resulted in an annual yield of 16.3% in case of *Strategy A*.

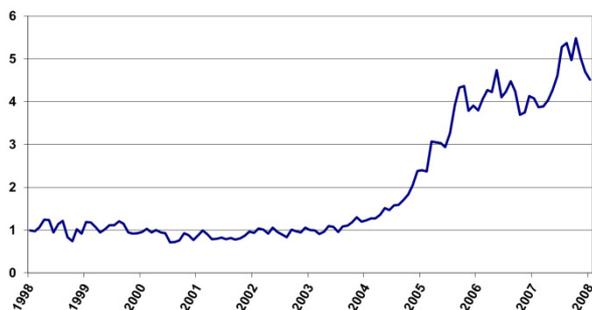


Figure 1: Price dynamics of MOL shares at BSE (monthly closing price data between 1998-2008, beginning of 1998=100%)

Source of data: Budapest Stock Exchange (www.bse.hu)

Now let us assume that *Strategy B* is available at an interest rate $r=12.0\%$. Should we take a loan at 12.0 per cent if you can invest the money in a stock yielding 16.3%? The answer, at least at first glance, seems to be simple but it is not. Do not forget that you have to repay the loan by selling the stocks. This redemption profile makes the strategy path dependent: it is not only the final stock price, but also the interim dynamics that count.

Let us check how *Strategy B* performs if $D=2$. Since the stock price starts from 1, *Strategy A* buys 1, while *Strategy B* purchases 3 units of stocks. After 120 months, *Strategy A* still has 1 stock and the value of the position is 4.5. However, the balance of *Strategy B* is shocking: we have only 0.35 unit of stock, which means

that we have lost almost 90 percent of our start-up portfolio and underperformed *Strategy A* significantly.

Figure 2 shows the net position dynamics of the two strategies. (Net position is the value of the stocks owned minus the outstanding debt.) *Strategy B* is above *Strategy A* in a few months at the beginning, but after the second year it remains below it all the way along. Hence, it would not have been a good strategy to finance the stock investment with an annual growth rate of 16.3% from a loan at an interest rate of 12%. The reason is the path of the stock price: loan payments during the long horizontal travelling of the stock price consume too much of our portfolio. By the time price will start to increase sharply, we have already sold approximately 75 percent of our original stock reserve. It is also notable that the net equity value of *Strategy B* goes negative several times. In these periods outstanding debt is higher than the value of assets. Although we allowed for this in our model, in the real life typically there is a minimum (positive amount) requirement for the shareholder's capital. (E.g. if that is not met for a period of time the firm has to be liquidated under the Hungarian law.) However, as long as the number of stocks owned is above zero, we still have some chance to repay the debt, if the stock price increases significantly (as in our case). The really serious problem occurs in case we would run out of stocks before repaying the total of the debt. The number of stocks cannot increase in *Strategy B*, hence once we have sold all of them, the firm goes bankrupt for sure.

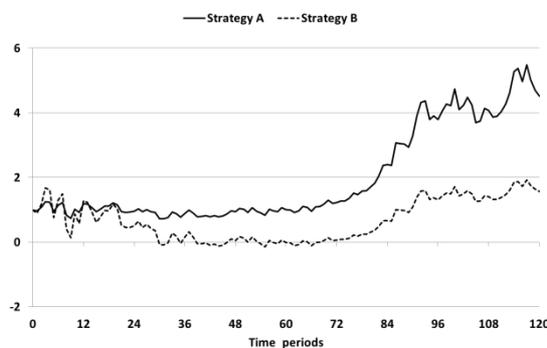


Figure 2: Net equity value of the two strategies ($D=2$ and $r=12\%$ in *Strategy B*)

In the rest of the paper, we seek the answer to the following questions:

- What is the *break-even interest rate* of the loan that makes the final net equity positions equal in the two strategies?
- How does the break-even interest rate depend on the size of leverage?
- What happens if we keep the initial and final stock prices, but change the path between them? How will this affect the performance and the break-even interest rate of *Strategy B*?

Break-even interest rate

Since the 12 per cent interest rate in the first numerical example was too high to make *Strategy B* profitable, the break-even interest rate must be under this level for sure. However, the exact answer was a bit surprising even for the authors: only interest rates under 5.11% could have been accepted so that we still have one stock at the end (that is, the two strategies end up with the same net position).

Table 1 collects the ultimate number of stocks in *Strategy B* for different leverage and interest rate values. The break-even interest rates are those with exactly 1 remaining stock. Some basic features of the data in Table 1 are as anticipated. For example, the higher the interest rate at a given leverage level, the less stock remains under *Strategy B*. Visually, the values are descending from the top to the bottom

An interesting result is the shaded row which shows that in case of 5.11% interest rate *Strategy B* will always have exactly 1 stock at the end. With other words, *the break-even interest rate is independent of the leverage*. In what follows, we show that this relationship is general and not the result of our specific data. Formalising the problem we have the following equations.

Periodic principal payment of the loan:

$$P_t = P = \frac{D_0}{nT} \tag{1}$$

Outstanding principal of the loan:

$$D_t = D_{t-1} - P_t = D_{t-1} - \frac{D_0}{nT} = \frac{nT-t}{nT} D_0 \tag{2}$$

Periodic interest payment on the loan:

$$I_t = \frac{r}{n} D_{t-1} = \frac{r}{n} \frac{nT-t+1}{nT} D_0 \tag{3}$$

Number of shares in *Strategy A*:

$$N_{t,A} = N_A = \frac{E_0}{S_0} \tag{4}$$

Number of shares in *Strategy B*:

$$N_{t,B} = \frac{E_0 + D_0}{S_0} - \sum_{i=1}^{nT} \frac{P_i + I_i}{S_i} \tag{5}$$

where r is the annual interest rate, T is the number of years examined, n is the number of payment periods within a year, $t=0, 1, \dots, nT$ is the index of time periods, and S is the price of the stock.

Our goal is to find the interest rate that makes the two strategies indifferent, hence where $N_{nT,A} = N_{nT,B}$. Substituting the previous equations will yield the following formula for r :

$$\begin{aligned} \frac{E_0}{S_0} &= \frac{E_0 + D_0}{S_0} - \sum_{i=1}^{nT} \frac{P_i + I_i}{S_i} \\ \frac{D_0}{S_0} &= \sum_{i=1}^{nT} \frac{P_i + I_i}{S_i} \\ \frac{D_0}{S_0} &= \sum_{i=1}^{nT} \frac{D_0 + \left(\frac{r}{n}\right)(nT - i + 1)D_0}{nT S_i} \\ \frac{nT}{S_0} &= \sum_{i=1}^{nT} \frac{1 + \left(\frac{r}{n}\right)(nT - i + 1)}{S_i} \\ r &= n \frac{\frac{nT}{S_0} \sum_{i=1}^{nT} \frac{1}{S_i}}{(nT+1) \sum_{i=1}^{nT} \frac{1}{S_i} - \sum_{i=1}^{nT} \frac{i}{S_i}} \end{aligned} \tag{6}$$

Table 1: Number of remaining stocks in *Strategy B*

		Leverage (V/E)											
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
Interest rate	0%	1.060	1.121	1.181	1.241	1.302	1.362	1.422	1.483	1.543	1.603	1.664	1.724
	1%	1.049	1.097	1.146	1.194	1.243	1.291	1.340	1.388	1.437	1.485	1.534	1.582
	2%	1.037	1.073	1.110	1.147	1.184	1.220	1.257	1.294	1.330	1.367	1.404	1.440
	3%	1.025	1.050	1.075	1.100	1.124	1.149	1.174	1.199	1.224	1.249	1.274	1.299
	4%	1.013	1.026	1.039	1.052	1.065	1.078	1.091	1.105	1.118	1.131	1.144	1.157
	5%	1.001	1.003	1.004	1.005	1.006	1.008	1.009	1.010	1.011	1.013	1.014	1.015
	5.11%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	6%	0.989	0.979	0.968	0.958	0.947	0.937	0.926	0.915	0.905	0.894	0.884	0.873
	7%	0.978	0.955	0.933	0.910	0.888	0.866	0.843	0.821	0.799	0.776	0.754	0.731
	8%	0.966	0.932	0.897	0.863	0.829	0.795	0.761	0.726	0.692	0.658	0.624	0.590
	9%	0.954	0.908	0.862	0.816	0.770	0.724	0.678	0.632	0.586	0.540	0.494	0.448
	10%	0.942	0.884	0.826	0.769	0.711	0.653	0.595	0.537	0.479	0.422	0.364	0.306
	11%	0.930	0.861	0.791	0.721	0.652	0.582	0.512	0.443	0.373	0.303	0.234	0.164
12%	0.919	0.837	0.756	0.674	0.593	0.511	0.430	0.348	0.267	0.185	0.104	0.022	

Since D does not appear in the last formula, the break-even interest rate is indeed independent from the leverage. Rather it depends on the length of the total investment T , the frequency of cash payments needed n and the path of the share prices S_t .

However, it is important to emphasize that the leverage is not irrelevant for the final net position of Strategy B. Higher leverage increases our profit in case r is lower than the break-even level, and decreases it if r is above the determined level.

The intuition behind the irrelevance of D to the break-even interest rate is the following. The break-even interest rate is the interest rate where D is neutral (the two strategies result the same final position), hence its level does not count in this problem. An in-depth understanding might be achieved if we imagine a firm financed only from one source (D) for which the cost of capital needs to be paid monthly. With a given price development path we can determine whether it is worth investing in the share. Our final decision would be independent of the actual quantity of capital as both return and cost of capital would be proportionate to the investment. Whether it is worth taking this capital or not is also independent of whether we have some extra capital (E) without regular cost of capital to pay. The leverage itself would only determine the extent of the effect, the direction of the effect is derived from the comparison of the path (series of returns) and the cost of capital.

PATH DEPENDENCY

We have already observed that the performance of *Strategy B* is path dependent: it is not enough to know that the price of the stock increases by 16.3% annually; it is also important how this average growth rate is achieved. Path dependency is a well-known term in connection with derivative products. American, Asian or other exotic options are path dependent since their payoff (and consequently their price) depends on the dynamics of the underlying product's price.

A great description of path dependent derivatives might be found in Wilmott (2006). Some aspects of the link between price dynamics and optimal hedging is described by Dömötör (2012, pp. 73-81.) for currency futures. Ruttiens (2013) describes the same path dependency problem when arguing that volatility is not always a good measure of risk.

Now we will illustrate how path dependency influences a simple leveraged stock investment, namely *Strategy B*. We will not deal with all the possible or probable ways that the stock price could have taken between the initial 1 and the final 4.5 values. (The number of such paths is of course infinite.) We are only examining paths with the same monthly logreturns as those actually occurred.

We determine the monthly logreturns (y_t) of the stock in the common way:

$$y_t = \ln \left(\frac{S_t}{S_{t-1}} \right) \quad (7)$$

Basic descriptive statistics of the monthly logreturns in our sample are summarized in Table 2.

Table 2: Descriptive statistics of the logreturns

Minimum	-38.01%
Maximum	32.01%
Mean	1.26%
Standard deviation	10.57%

Extreme paths

We will show the significance of path dependency by changing the order of occurrence of the logreturns. First, we discuss two extreme trajectories, and sort the actual logreturns in ascending and descending order. Figure 3 shows the actual and the two extreme paths of the stock price. Since the initial value and the set of monthly logreturns are the same in all the three cases, the stock price will always be 4.5 at the end and the average annual growth rate is also unchanged (16.3%). However, rearranging the logreturns causes massive variations in the paths, so much so that we had to apply different axis-scaling for the descending scenario.

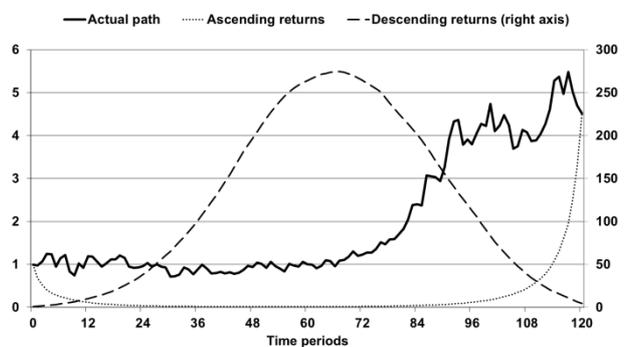


Figure 3: Paths of the stock price with actual and ascending (left axis), and descending (right axis) order of given logreturns

It is trivial that with given logreturns the descending scenario is the best, while the ascending one is the worst case for *Strategy B*. In these extreme scenarios, with $D=2$ and $r=12\%$, the leveraged strategy ends up with 2.82 and -88.12 stocks respectively. The negative sign of the second data shows that in the worst case scenario we run out of stocks and cannot repay the loan. In other words we go bankrupt before the end of the investment period.

Actually, we go bankrupt very quickly: we start with 3 stocks, but we are not able to finance the loan payments through more than 14 periods. That is, in almost one year we have to sell all of our stocks and we still have approximately 1.75 debt to repay. Since the returns are in ascending order in this case, big negative returns come first in the early periods, and by the time better (positive) returns would occur we will have already no stocks and cannot benefit from the favourable price movements.

The significance of the path is even more obvious if we determine the break-even interest rates in the two extreme scenarios: it is 270% in case of descending returns and -17% with ascending returns. The later means there is no realistic macroeconomic situation (r) where leverage would pay off unless there would be a regular cash inflow from our investment, for example we could collect significant dividend payments.

Random paths

After the two extreme scenarios we examine how random occurrence of the logreturns would perform. We simulated 2,000 scenarios by combining the given logreturns randomly. Figure 4 shows the actual and three simulated paths. The initial and the final stock prices are still 1 and 4.5 respectively; since logreturns are additive and changing the order does not affect the sum.

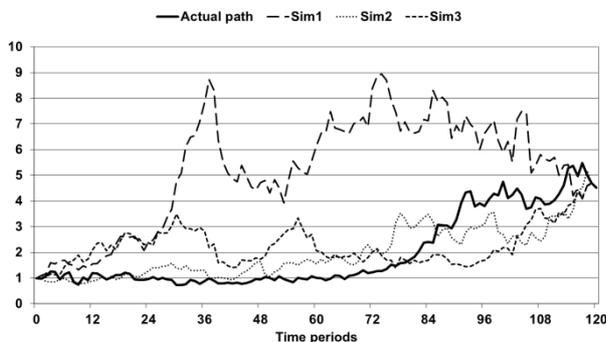


Figure 4: Actual and simulated paths with given set of logreturns

We have also calculated the break-even interest rates and the final number of stocks in *Strategy B* with the previous parameter settings ($D=2$ and $r=12\%$). We also quantified the amount of cash inflow (e.g. dividend, capital increase by shareholders, fresh debt) needed to save a path generating bankruptcy (negative number of shares). Table 3 summarizes the descriptive statistics of these calculations, while Figure 5 and Figure 6 show the relative frequencies of the simulated data.

Out of the 2,000 scenarios simulated, the maximum number of shares is 2.28 which means that the maximum final net position is as high as 10.26 ($2.28 \cdot 4.5$). Since we achieved it with 1 unit of equity,

this growth equals to 1,026% total return over the 10-year investment horizon. Similar calculations show that the average total return is 464%, while the median is 527%.

Table 3: Descriptive statistics of the break-even interest rate, the number of stocks and the cash needed to save in *Strategy B* ($D=2$, $r=12\%$)

	Break-even interest rate	Number of remaining stocks	Cash needed to save
Minimum	-9.59%	-5.32	-2.08
Maximum	64.55%	2.28	0.00
Mean	15.85%	1.03	-0.08
Std. dev.	10.57%	0.72	0.29
Median	14.61%	1.17	0.00
1 st quartile	8.06%	0.67	0.00
1 st decile	3.59%	0.13	0.00

These figures are equivalent to 26.2% (maximum), 16.6% (mean) and 18.1% (median) annual returns. We may observe that the average annual return is very close to that of *Strategy A* (16.3%).

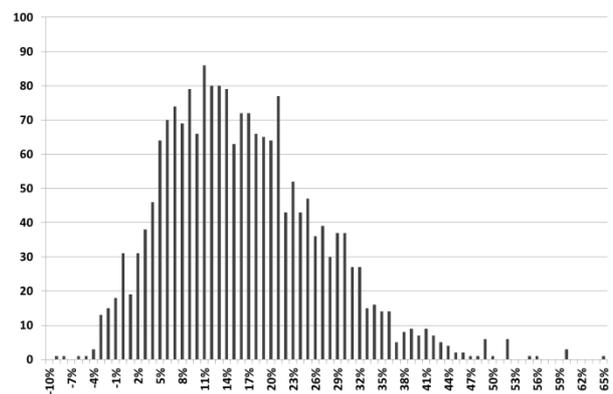


Figure 5: Frequencies of break-even interest rates from 2,000 simulated paths

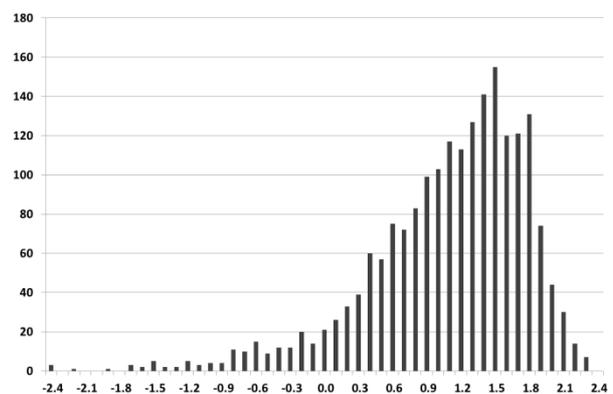


Figure 6: Frequencies of the final number of stocks in *Strategy B* from 2,000 simulated paths

Although our average annual return on the stock investment (*Strategy A*) is 16.3%, in 56.15% of the cases a similar cost for financing would have decreased our returns. In half of the paths an interest rate below 14.61% would be required, and we need to push down interest until 8.06% to have 75% chance to boost our profit by using the given leverage. For a 90% chance to win on the debt one needs a rate below 3.59%. Even if debt is for free we have just 95.8% chance that we will be able to cover all principal payments on time.

Using our initial interest rate of 12%, only 59.7% of all paths make leverage more profitable than *Strategy A* with no debt. Out of 2,000 only 1,841 (92.05%) scenarios will end up with positive number of shares, this means that the probability of bankruptcy is 7.95%. Given the possibility to survive bankruptcy by using fresh capital the total amount of money needed to save our company in these 159 cases is shown in Figure 7. (No cost for capital assumed.) To save our firm in all cases we would need 2.08 unit of money that is more than the double of the initial equity capital (E) available. To push the probability of bankruptcy down to 5% (that is to save 2.95% of the paths) we need 70% more start-up equity kept in form of cash, while an extra cash reserve equal to original E would reduce chance of failure to 3.9%.

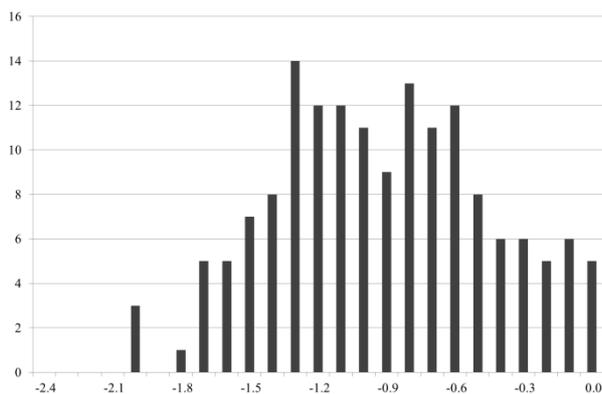


Figure 7: Frequencies of fresh capital needed to save 159 bankruptcy-generating simulated paths

CONCLUSION

In this paper we investigated path dependency in case of an investment strategy using no path dependent investment assets. We illustrated how the use of a path dependent “negative” asset (debt) would make our strategy path dependent. This kind of risk is very often overlooked by investors making them unable to correctly judge total risk taken.

We have shown that the break-even interest rate for using debt in financing any portfolio is independent of the leverage itself. Although, this well-known general

rule in financing is dramatically different in case of path dependency: the critical rate is not necessarily equal to the return that can be achieved by investing the capital, rather than it may be considerably lower or higher.

The risk path dependency generates is not negligible. In case of our random simulation in 7.95% of the cases the firm went bankrupt while earning an average asset return higher than the interest rate of debt over the total investment period. These failed firms would need serious amount of fresh capital to survive but even then most of those would produce loss.

Our research has also raised several further questions. For the debt different repayment schedules could be considered, fresh capital may have some alternative cost, and also length of surviving could be measured at different leverage levels. Further, our model could be developed by requiring a certain level of equity for allowing a company to move on to the next period of time.

REFERENCES

- Adner, R. – Levinthal, D. A. 2004. "What is not a real option: considering boundaries for the application of real options to business strategy" *Academy of Management Review*, Vol. 29, Issue 1, 74.
- Bianco, M. – Gerali, A. – Massaro, R. 1997. "Financial systems across “developed economies”: convergence or path dependence?" *Research in Economics*, Vol. 51, Issue 3, September 1997, 303-331.
- Dobusch, L. – Kapeller, J. 2013. "Breaking new paths: theory and method in path dependence research." *Schmalenbach Business Review (SBR)*, Vol. 65, No. 3, 288-311.
- Dömötör, B. M. 2012. A finanszírozási likviditás hatása a piaci kockázatok kezelésére, PhD dissertation proposal, Corvinus University of Budapest
- Graves, W. 2011. "The Southern Culture of Risk Capital The Path Dependence of Entrepreneurial Finance." *Southeastern Geographer*, Vol. 51, No. 1, 49-68.
- Jazaerli, S. – Saporito, Y. F. 2013. "Functional Itô Calculus, Path-dependence and the Computation of Greeks." arXiv
- Pierson, P. 2000. "Increasing Returns, Path Dependence, and the Study of Politics" *The American Political Science Review* Vol. 94, No. 2 (Jun., 2000), 251-267.
- Ruttiens, A. 2013. "Portfolio Risk Measures: The Time's Arrow Matters." *Computational Economics*, Vol. 41, No. 3, 407-424.
- Száz, J. 2013. "Meddig ér a takaró?" *Hitelintézet Szemle*, Vol. 12, No. 6, 458-468.
- Thompson, A. C. 1995. "Valuation of Path-Dependent Contingent Claims with Multiple Exercise Decisions over Time: The Case of Take-or-Pay." *Journal Of Financial & Quantitative Analysis* 30, No. 2, 271-293.
- Wilmott, P. 2006. *Paul Wilmott on Quantitative Finance*. John Wiley & Sons, Chichester. 2nd edition.

AUTHOR BIOGRAPHIES

Ágnes VIDOVICS-DANCS is assistant professor at the Department of Finance at Corvinus University of Budapest. Her main research areas are government debt management in general and especially sovereign crises and defaults. She worked as a junior risk manager in the Hungarian Government Debt Management Agency in 2005-2006. Her e-mail address is: agnes.dancs@uni-corvinus.hu

Péter JUHÁSZ, PhD, CFA is associate professor at the Department of Finance at Corvinus University of Budapest. Besides teaching he also works as a trainer and management consultant mainly for SMEs and governmental entities. His field of research covers business planning and valuation, corporate risk management and VBA programming of MS Excel. His e-mail address is: peter.juhasz@uni-corvinus.hu

János SZÁZ, CSc is full professor at the Department of Finance at Corvinus University of Budapest. He is the first academic director and current president of the International Training Center for Bankers in Budapest. Formerly he was the dean of the Faculty of Economics at Corvinus University of Budapest and President of the Budapest Stock Exchange. Currently his main field of research is financing corporate growth when interest rates are stochastic. His e-mail address is: janos.szaz@uni-corvinus.hu

MODELLING THE COLLAPSE OF A CRIMINAL NETWORK

Martin Neumann
Ulf Lotzmann

Institute of Information Systems Research
University of Koblenz-Landau
Universitätsstraße 1, Koblenz 56070, Germany
E-mail: {maneumann, ulf}@uni-koblenz.de

KEYWORDS

Qualitative data analysis, conceptual modelling, agent-based simulation, normative agent architecture, criminal networks.

ABSTRACT

Research activities aimed to understand the dynamics of criminal networks or organisations – like extortion racket systems – are of interest for criminologists and scientists in related fields, for practitioners from police and judiciary, as well as for political decision makers. This paper aims to contribute to this area by bringing together data analysis methods with conceptual modelling and simulation techniques employing normative agents. As an exemplary case the internal collapse of a criminal network is investigated. The outlined model development process involves qualitative data analysis of police files, specification of a conceptual model of the dynamics within the regarded criminal network, and the definition of intra-agent processes for the agents used in the simulation model.

INTRODUCTION

The research outlined in the paper is part of the GLODERS research project (<http://www.gloders.eu/>), directed towards development of an ICT model for understanding the dynamics of Extortion Racket Systems (ERSs). These are criminal organisations of which the Mafia is but one example. Two scenarios have been selected to develop conceptual models in complementary ways which then can be transformed into formal simulation models:

- A Palermo scenario, describing the dynamics of the (at least formerly) culturally deep entrenched Sicilian Mafia within the society. This theory driven scenario can build on rich empirical research about the Sicilian Mafia, the Cosa Nostra.
- A contrasting scenario, describing the internal dynamics within a criminal organisation in a different European country. This data driven scenario applies a grounded theory approach, building on Police interrogations resulting from a number of investigations of a particular criminal group. This scenario is subject of this paper.

Internal dynamics of criminal organisations is of theoretical interest as subject of scientific research because criminal organisations face the specific problem of securing compliance with the norms of conduct in the absence of juridical norm enforcement. This generates a problem of conflict resolution: Theories of norms typically refer to the

notion of sanctions or punishment as a norm enforcement mechanism (Ullman-Margalit 1978, Hechter and Opp 2001, Bicchieri 2006, Horne 2007). In behavioural terms punishment can be described as some kind of aggression. If I get a fine for wrong parking, or if a mother scolds her child, this goes along with losses of utilitarian values (anyhow they are measured: in the case of a fine it is monetary value, in the case of scolding it might be the feeling of shame). However, obviously not every aggression is punishment. This is particular obvious in the case of criminals. Striking an old lady down to get her purse is aggression but certainly it is a norm deviation, not enforcement. Likewise extortion is a crime. However, without recourse to state monopoly of legitimate violence the motivation of aggression might remain ambiguous. For instance aggression can be self-interested. Robbery of a purse is one example. As another possibility it might simply be result of extensive drug consumption. In the case of criminal organisations this ambiguity is particular precarious. On the one hand, there is a need to enforce the norms of conduct in some way. On the other hand, the lack of the authority of juridical court provides an incentive to manipulate norm enforcement in its own interest. For instance Arlacchi (1992) describes the situation of the Cosa Nostra in the 2nd Mafia war in the 1980s (Arlacchi 1992, Dickie 2004) as a Hobbesian society of a war of everybody against everybody else. While formally aggression has been justified as sanction it became obvious that this justification was merely cheating. Thus aggression remains ambiguous. In the legal society recourse to the jurisdiction of the court provides an ultimate answer. This is absent in criminal organisations. For this reason criminal organisations provide a means to investigate mechanisms for conflict resolutions and how these might fail. In particular this involves reasoning about aggression (Andrighetto et al. 2013). This is a particular challenge for simulation of normative processes in which punishment is taken as a basic concept and thus the need to interpret aggression is not considered so far (Axelrod 1986; Conte and Castelfranchi 1995; Andrighetto et al 2007; Neumann 2008; Savarimuthu et al. 2013; Lotzmann et al. 2013).

This paper focuses on the scenario for which the data-driven modelling approach has been applied in order to find the relevant aspects of the dynamics that led to the breakdown on this particular criminal network. There are annotations from the real case used to show traceability of the conceptual model to the evidence base. However, the police files that constitute the evidence base for this scenario are not publicly available, thus the description had

to be rephrased in a neutral way in order to ensure protection of privacy.

EVIDENCE BASE OF THE MODEL

The Scenario applies a grounded theory approach (Corbin and Strauss 2008) based on police interrogations in 2005 and 2006 that resulted from various police investigations of a criminal gang. Established in the early 1990s its business model consisted of drug trafficking and laundering the illegal money gained in the drug business. Drug trafficking was done by 'black collar criminals' with access to the production and distribution of drugs. 'White collar criminals' were ordinary businessmen responsible for the money laundering. They got roped into the business in the early 1990s. The psychological techniques applied to draw them in the illegal world beyond a point of no return will not be subject of this paper. Police files identified (at least) one white collar criminal working in the real estate business. It is important that the real estate trader had a good reputation in the legal society. This allowed him to invest illegal money in the legal market and give the return of investment back to the investor. Money laundering is essentially based on a norm of trust: the black collar criminals need to hand over the money to their partners and trust them that they will get the return of investment back from the trustee. In a covert organisation this cannot be secured by formal contracts. Therefore trust is essential. The network lasted for about 10 to 15 years until it collapsed. An initial divide went out of control, and the mistrust could not be encapsulated but spread rapidly through the whole network. Once trust was corrupted, a run on the bank was initiated. Black collar criminals attempted to get their invested money back before it was lost completely. Attempts to get the money back led to extortion. Thereby the white collar criminal became victim of his criminal business partners. A formerly symbiotic relationship between black and white collar criminals (a long term relation of a win-win situation for both) became parasitic (i.e. a lasting but no longer profitable situation). This generated a cascading effect through the network which destroyed the overall network in a violent blow-up. This characteristic of the case makes the data particularly interesting to identify essential elements in the mechanisms of conflict resolution in the absence of juridical law, i.e. the failure allows to identifying the elements which must not be missing.

Conflicts escalated to a degree of violence that has been described by witnesses as a 'rule of terror' in which 'old friends were killing each other'. In fact, many members of the network were killed. The 'rule of terror' could not be attributed to an individual member of the group but can be described as ruled by an invisible hand. Norms of conduct were only implicit, leading to many misperceptions which generated a cascading effect of spreading mistrust. This shall be illustrated here by one example: In the data it can be found the testimony of a member that "M. told the newspapers [about my role in the network] because he thought that I wanted to kill him to get the money." M. had survived an attack on his life, but he was wrong in the assumption that this particular member of the organisation was behind this attack. However, his counter-reaction caused further panic of other group member such as the one

who had been reported here and wrongly brought into trouble. This was the starting point of the cascading of mistrust in the group.

This example provides insights into processes of reasoning about aggression: first, M. interpreted the attack on his life not as a penalty (i.e. death-penalty) for deviant behaviour from his side such as being too greedy. Instead he concluded that the cause of the attack was based on self-interest (the other criminal 'wanted his money'). Thus he interpreted the attack as norm deviation rather than enforcement. Next, he attributed the aggression to an individual person and started a counter-reaction against this particular person by betraying 'his role in the network'. However, since his reasoning went wrong, this counter-reaction provoked further reasoning about the cause of and possible reactions to his aggression. This generated a cycle of revenge and counter-revenge.

In more abstract terms the cascading effect of spreading mistrust is due to the fact that the criminal group was based on personal acquaintanceship without a formal structure of positions. For this reason it remained precarious to differentiate between punishment and revenge, depending on subjective interpretation which factually initiated a lot of misunderstanding. Punishment entails a stop point for the aggression: Punishment is applied for a particular situation, e.g. a fine for wrong parking. Once the fine is paid, the punishment is over and the aggression terminates. As this example demonstrates, revenge might lead to counter revenge which again stimulates new aggression. However, interpretation of aggression as punishment requires at least minimal social structure: aggression applied by a higher hierarchy level is more likely to be interpreted as punishment (at least members may obey even if factually it might be wrong), whereas aggression between peers is more likely to be interpreted as revenge.

METHODOLOGICAL APPROACH

Methodologically the data was loaded into MaxQDA as a tool for qualitative text analysis (see Corbin and Strauss 2008) and text passages were annotated which then were summarised into codes deriving concepts from data. Concepts stand for classes of object, events or actions which have some major properties in common. This enables concept identification supported by CAQDAS technology. The coding derived with MaxQDA served as the basis for concept relation identification with the CCD tool (a software for creating Consistent Conceptual Descriptions; Scherer et al. 2013). The CCD tool provides an environment for developing a conceptual model by a controlled identification of condition-action sequences (denoted as action diagram) which represent the micro-mechanisms at work in the processes described in the data. Whereas the data describes individual instantiations, the condition-action sequences represent mechanisms insofar as they describe generalisable event classes. However, empirical traceability is ensured by tracing the individual elements of the action diagram resulting from the identification of condition-action sequences in the CCD tool back to text annotation in the data. These annotations are extracted from the coding derived with MaxQDA. This approach is particular appropriate for dissecting cognitive elements in the data. Police interrogations are a situation of

dialogical conversation, not biased by categories in the mind of the researcher. This allows to bringing the empirical analysis very close to the subjective perception of the actors. An in-depth analysis of subjective meaning attributed to certain situations enables to establish an empirical evidence base for modelling intra-agent processes such as reasoning about aggression.

The agent architecture constituted by these intra-agent processes is the starting point for creating several types of software agents, which are able to interact within a simulation environment by different means, as described in the evidence base and specified in the conceptual model.

CONCEPTUAL MODEL OF THE COLLAPSE OF THE GROUP

In the following the action diagram resulting from the data analysis will be exemplified. In the process of the collapse of the group four phases can be distinguished: the ordinary business of money laundering, a crystallising kernel of mistrust, a stage of conflict escalation and finally the 'corrupt chaos', including a run on the bank. The first two processes will be illustrated in detail, the others will only briefly be sketched.

The ordinary business: money laundering

The ordinary process of money laundering (Figure 1) starts with illegal money available and ends when legal money is available for the black collar criminals. In the following this process is shown in detail.



Figure 1: Ordinary business of money laundering

In this diagram the conditions of, for instance, 'level of trust above threshold' and 'illegal money available' trigger the action 'give money to trustee'. Traceability of the empirical evidence is provided by annotations of the condition-action sequences. In the following, passages of the police interrogations will be documented which provide

the empirical evidence for all phases of the process. Names and dates have been hidden for reasons of protecting privacy.

Annotation (illegal money available): "In the period between 1990 and XX 1992 police investigations had been undertaken. These revealed a criminal organisation concerned with drug trafficking. The report from XX 1992 estimated the income and the costs. It is estimated a transaction volume of nearly 300 million."

If illegal money is available a second condition of trust need to be fulfilled for starting the process.

Annotation (level of trust above threshold): "O1 and V01 seem to be friends for me."

These annotations secure traceability of the starting conditions for the process. Money laundering is triggered when illegal money is given to a trustee with a legal business who invests the money in the legal market. The trustee is the link between the illegal and the legal world. The next step in the process is the investment of the money in the legal market as demonstrated in the following annotation:

Annotation (return of investment available): "On the basis of witnesses and financial investigations it is suspected that O1 and all persons directly or indirectly associated with him received considerable boni for transactions in which V01 and his companies had been involved."

The return of investment is redistributed to the investors. However not directly but via straw men who receive the money from the trustee and hands it over to the original investor.

Annotation (straw man received money): "The funding went from V1 to B3 and then to the father of M.O."

This annotation shows the process of money flow via so-called straw men. Finally legal money is available for black collar criminals.

Annotation (legal money available): "At the moment I have paid 800 000 in the firm which are now several millions worth through legal trade."

Crystallising kernel of mistrust

The action diagram in Figure 2 shows the details of the initialisation of mistrust, followed by annotations, demonstrating the empirical traceability of the condition-action sequences.

Starting point is the event that for some reasons (out of the scope of the investigation) some member of the organisation becomes distrusted, as illustrated by the following annotation.

Annotation (member X becomes disreputable): "Since O8 is released from prison there were tensions observable between O6, O1 and V01 on the one hand and O8 on the other hand."

This triggers an aggressive action against this member, who needs to interpret the aggression once he recognises it. First an example for aggression will be provided.

Annotation (perform aggressive action against member X): "An attack to the life of M."

Aggression may be interpreted either as norm enforcement, i.e. a form of punishment for deviant behaviour (Norm of trust demanded), or as an arbitrary aggression (Norm of trust violated). In the former case the victim of aggression may obey, which restores the trust in the organisation, or

cheat. In the latter case the victim of the aggression decides about the reaction by either betraying the organisation or performing an act of counter-aggression. This is demonstrated by the examples of empirical evidence below.

Annotation (member X decides to betray criminal organisation): Statement of V01: "M. told the newspapers 'about my role in the network' because he thought that I wanted to kill him to get the money."

As already indicated M.'s interpretation of the aggression was misleading. However, it is an example that he interpreted the aggression as a violation of his trust in V01 and reacted by betraying him. An example of counter-aggression is the following:

Annotation (member X performs counter-aggression): "Presumably V01 asked the Hells Angels to make an operation against O1 in return for a huge amount of money."

A different example is the following sequence, showing the pathway to obedience is the following sequence of aggression and the corresponding reaction.

Annotation (aggression recognised by member X): "At May XX, YY, O5 came to my house in order to say that at 8 in the evening I should come to the forest. This is standard: intimidate and request for money."

Annotation (member X obeys): "I paid, but I'm alive."

However, the victim of aggression may also decide to cheat:

Annotation (member X decides to cheat): "Following O1 C. betrayed to him even the people who they wanted to liquidate."

Conflict escalation

The next sequences will only briefly be sketched. Selected annotations indicate the empirical evidence. The process of conflict escalation is an expansion of the initial mistrust. In particular a feedback is included from 'becoming victim of aggression' to 'interpret aggression'. In the case of counter-aggression performed by the original victim of an aggressive act, a new member of the group becomes victim of aggression, who in turn needs to interpret this aggression. This enfolds a positive feedback loop which may become unstable.

Betrayal can appear in various forms: whereas internal betrayal means that the person does something incorrect against a suspected aggressor, external betrayal consist of whistle-blowing to the police or even the wider public (such as e.g. a newspaper). Internal betrayal leads to the event that this member of the organisation becomes disreputable only if this becomes known.

Annotation: "V01, killed at May XXX had spoken several times with criminal investigation officers." The text example above provides also an example for the escalation of violence up to murder. In fact several members of the group were killed.

If the organisation becomes public (by external betrayal or because of visible violence such as murder), it becomes possible to fight against the criminal organisation from outside with police investigations. These might lead to juridical decisions. Both events might trigger a counter reaction which will be explained in the next section.

'A corrupt chaos'

Finally the escalation of the conflict may reach a stage at which trust is no longer recoverable. In particular this includes the cognitive element of 'fear for life' and the modelling perspective on this subjective cognition of 'evaluating the level of trust'.

The process of the collapse of the organisation can be described as a cascading effect in which a norm of trust in the organisation breaks down. A subjective perception of the overall situation is given below:

Annotation: "There is a rule of terror in the town."

This overall situation consists of several micro-elements. However, these were not visible for all members of the group, leading to the nebulous assumption of a rule of terror which could not be attributed to a single person. Subjectively the terror seemed to be ruled by an invisible hand.

Annotation: "There is a corrupt chaos behind it."

In modelling terms, this can be described as result of an evaluation of the trust. If the trust in the organisation still remains above a threshold, the usual business continues. If this is not the case it might trigger a panic reaction. Moreover, group members may also become victim of

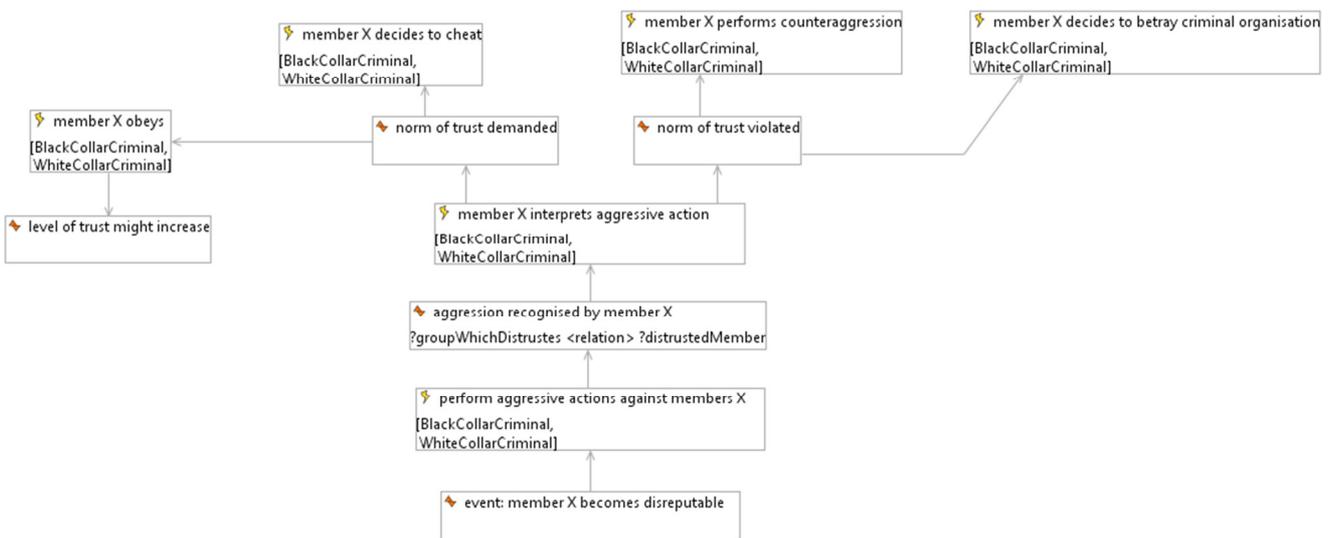


Figure 2: Crystallising kernel of mistrust

murder. The observation of the murder might again trigger the afore-mentioned panic reaction. A panic reaction is characterised by immediate and often irrational actions guided by fear. However, the object of a panic can be twofold: fear for life or fear for money. It remains unclear in the data if both are inherently intertwined or if they – at least possibly – might be distinct instances of panic. It will be a key objective of simulation experiments to investigate scenarios which clarify the causal relations between both kinds of panic. First, fear for life shall be considered. Fear for life triggers various possible reactions such as attempts of pre-emptive murder. This is illustrated in the following testimony:

Annotation: “He was at a point in which he was in a totally despaired situation. HLJ had several times tried to counteract. He had a plan to approach O1 with a weapon. However, in the last moment he didn't dare. At a different time he had two pistols with him. He planned to shoot O1 to death and to pass the other weapon in his hand in order that it appeared as if he had shot in self-defence.”

Run on the bank

The ‘corrupt chaos’ sketched above triggers a ‘run on the bank’. Fear for money stimulates a request to get the invested money back before it is lost completely. An attempt to get the money back nevertheless might trigger intimidation of the trustee (the white collar criminal) who now becomes victim of aggression which he needs to interpret; i.e. this enfolds a positive feedback cycle.

Annotation: “In the last year he was strongly under pressure because he had been extorted. That's what he said to me.” However, if the intimidation of the trustee becomes public in the organisation, this might trigger a run on the bank by stimulating a panic reaction among other group members. This is illustrated by the example below.

Annotation: “Soon after his death the widow of K had an affair with O1. She extorted 7 million from V01. Contrary to the claim of M. his entitlements had not been captured by this deal.”

This provided an incentive for M. to try to get his money back on his own.

INTRA-AGENT PROCESSES

This section describes the first steps of the process of developing a simulation model from the conceptual description presented above. The two pilot scenarios which are developed are meant to integrate the normative approach based on EMIL-IA (the normative agent architecture developed in the EMIL project, extended by norm internalisation capabilities; Andrighetto et al. 2010) with properties of several stylised facts models.

The following subsections firstly give an overview on some important simulation model properties in order to exemplarily describe the formalisation approach for distinct behavioural aspects of individual agent types (as derived from the action diagram, which in contrast covers the behaviour of the entire system). The intra-agent processes are defined as modules and specified by flow charts, focussing on processing of data, in this context mainly events for triggering processes, and different kinds of parameters determining the control flow. The most

important kind of parameters is related to norms ruling the agents' behaviour.

Simulation model properties

The conceptual model as derived from the qualitative data analysis comprises four major types of actors, which will become agent types in the simulation model. Three of the actor types constitute the criminal network:

- The black collar criminal, who is involved in drug trafficking, applies violent actions in various situations and for several reasons and needs the service of white collar criminals for money laundering.
- The white collar criminal, who is key person for money laundering and usually is less involved in violent actions. However, in the scenario of breakdown of the criminal network, his involvement in violence (i.e. the adoption of behaviour of black collar criminals) seems essential for the process.
- The (so called) straw men are supporting the other types of criminals in the money laundering process in different fashions.

The only actor outside the criminal organisation regarded in this scenario is the police. The actions of the police actor are currently limited to a general criminal investigation (eventually leading to a juridical decision) and the possibility of information leakage to actors of the criminal network.

All these actors are ruled by norms. As a result of the detailed examination of the empirical data, a restricted number of norms have been identified which implicitly govern the behaviour of the actors. As an example, for all types of criminals a ‘top-level’ moral norm exists:

NORM(1) “moral norm”: NOT VIOLATE TRUST c o

where c is a criminal and o is the criminal organisation or network. This norm describes the commitment to the norm of trust within the organisation which holds in the case of unexpected events and is entangled with interpretation of aggressive actions, self-reflexion and the consideration of own past actions.

Related to this norm, a number of concrete obligations are defined. An example is

NORM(1.3) “obligation”: PUNISH c_1 c_j IF c_j VIOLATE NORM(1)

where c_1 is a criminal who punishes the deviant criminal c_j for a norm violation.

Such a punishment triggers a ‘reasoning on aggression’ process within the punished agent, where the agent must decide whether the experienced aggression was such a punishment, or rather a self-interested act of aggression. This process is detailed to some extent in the following description of the architecture of one of the agent types, the black collar criminal.

Agent architecture of the black collar criminal

The architecture of the black collar criminal agent is defined by a number of processes which it partly shares with the other types of criminal agents (Figure 3). One of

these processes – the normative process – is of particular importance, as this encapsulates the core of normative reasoning and provides the link – even on software engineering level – between the simulation models for the two scenarios.

The other scenario-dependent intra-agent processes follow in large part the conceptual model, i.e. the behaviour described in the CCD action diagram is reflected in these processes. Exemplarily, two of these processes should be outlined in some more depth: ‘Reasoning about aggression’ together with ‘Reacting on aggression’. Subsequently, the functional interactions between these two processes and the ‘Normative process’ are shown.

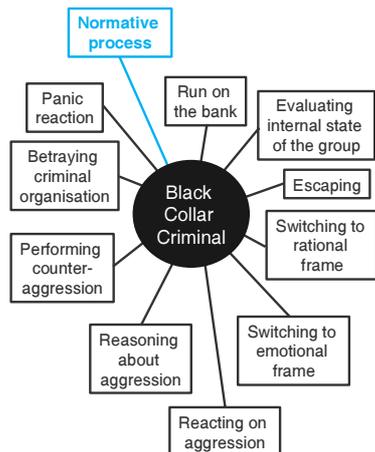


Figure 3: Intra-agent processes of the black collar criminal agent

Reasoning about aggression

The ‘Reasoning about aggression’ process (Figure 4) is triggered when the agent recognises an aggression against itself. It comprises the first of three stages of a decision process, eventually leading to possible reactions on the aggression.

In the first stage it is decided whether the aggressor is reputable and the motivation for the aggression is not gratuitous. Information on trustworthiness of the aggressor from an ‘image and reputation repository’ (a data structure which stores the agent’s belief on image and reputation of other fellow agents) is regarded here. If the aggressor is reputable, a possibly normatively motivated aggression is anticipated and the normative process is triggered at the second stage. A possible result of the normative process might be that the inherent sanction recognition failed (see subsequent section), but the aggression poses a potential threat to the agent. In this case, and in the case that the aggressor is recognised as not reputable, reactions will be triggered by entering the third stage of the process in which the operational mode of the agent is either set to a rational or an emotional frame, amongst others depending on the strength of the initial aggression.

The actual switching to one of the two frames is done in two separate processes not shown here, followed by triggering the ‘Reacting on aggression process’ (Figure 5), in which the agent decides how to retaliate the aggression (either by counter-aggression or by betrayal of the criminal network, depending on the mental frame which the agent has adopted before). This process can also come into play if the agent decides to cheat, i.e. a sanction is recognised

within the ‘Normative process’ but the agent decides not to obey the norm behind the sanction but rather to follow some other (individual) drives.

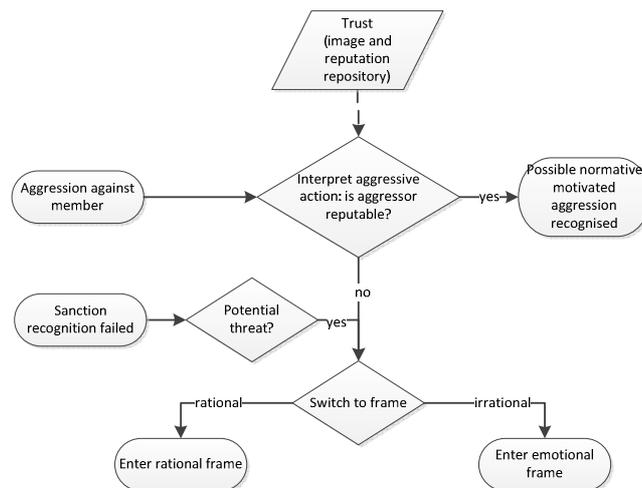


Figure 4: Intra-agent process for reasoning about aggression (rounded boxes are start and end events for the process, rhombi are decisions, parallelograms stand for parameters influencing decisions)

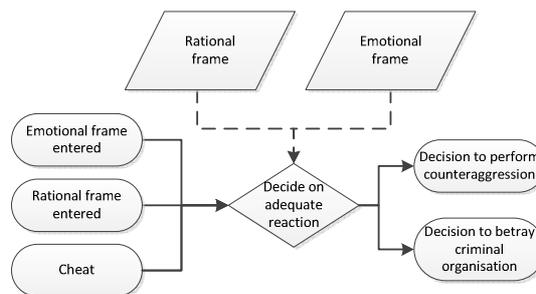


Figure 5: Intra-agent process for reacting on aggression

Normative process

The ‘Normative process’ is one of the major contributions of the GLODERS project and has been designed and is currently implemented with respect to support normative behaviour of agents in different kinds of simulation models. The architecture of this process is an extension and generalisation of EMIL-IA, which itself has been built upon EMIL-A, the normative architecture developed in the EMIL (Emergence in the Loop) project (Conte et al. 2014). The quite complex software design of this process contains six main components:

- An event classifier, pre-processing the incoming data (events like actions, sanctions etc., observation of such events, explicit norm invocations) and determining which of the following sub-processes to trigger.
- A ‘Norm Recognition’ sub-process, which tries to find a norm related to an action or sanction in the internal normative board (a database of norms, sanctions, relations between these two entities and salience values of particular norms).
- A ‘Norm Enforcement’ sub-process, which selects appropriate sanctions as reaction on perceived norm violations.
- A ‘Norm Salience’ sub-process which updates information on the salience of norms.

- A ‘Norm Adoption’ sub-process for normative learning.
- A ‘Norm Compliance’ function which calculates a normative drive, reflecting the strengths and salience of a norm. This value is used by several concrete decision processes of agents.

For the model presented in this paper, the normative process is triggered on two different events. Firstly, it is triggered by the observation that a member of the criminal network has become disreputable. This can be an external event, or due to (re-)actions in relation with the break-down of trust as specified by the conceptual model. This event is classified as observation of an (aggressive) action for which is checked whether a norm violation might be associated with the action. If this is the case, an appropriate sanction is issued.

The other triggering event of the normative process is associated with the ‘Reasoning about aggression’ process described in the previous section. If a possible normatively motivated aggression is recognised, the normative process then checks whether this aggression is a sanction or an aggression not related to any norm. In the former case, the agent can either obey to the norm, or intentionally cheat by not complying with the sanction related to a (potentially salient) norm and rather follow (even more salient) individual drives. For the latter case, a reaction on the aggression will be the consequence, expressed by the event ‘Sanction recognition failed’, signalled by the normative process.

The implementation of the sanction recognition sub-process is based on a search mechanism, trying to match the identifier of an experienced aggression (e.g. PUNISHMENT FOR TRUST VIOLATION) with an existing norm (in this artificial example the NORM(1) “moral norm”: NOT VIOLATE TRUST c o). These concepts have to be expressed in a formal language, which at the same time constitutes the interface to the ‘Normative process’ software component, i.e. which is used for defining the data stored within and processed by this component.

CONCLUSIONS AND FUTURE WORK

This paper shows the results of the conceptual modelling of the collapse of a criminal network. The qualitative data analysis informing the conceptual model as well as the first formalisation activities towards a simulation model are outlined with emphasis on important design details, e.g. the realisation of normative behaviour. The conceptual modelling enables dissecting the micro-mechanisms of a complex empirical process. These enable a certain degree of generalisation beyond a narrative story of a certain case to shed light on the wheels of social processes. Nevertheless, the evidence based modelling approach retains traceability of the abstract mechanisms to the empirical social world.

The model implementation phase has just started. In the current state of development, the most important behavioural aspects of the organisational breakdown have been regarded in a level of abstraction which makes the development of a sophisticated simulation model reasonable. However, some details and aspects (mainly regarding the ‘everyday business’ of the criminal network –

drug trafficking and money laundering) have not been treated in detail so far, as the data analysis of this part has not been completed yet. The most important short-term future task is to continue the incremental implementation of the concepts and processes described above. The simulation model will then contribute to computational normative agents by implementing reasoning about aggression whether or not to interpret it as sanction.

ACKNOWLEDGEMENTS

The research has been undertaken as part of the GLODERS project, co-funded by the European Commission under the 7th Framework Programme. This publication reflects the view only of the authors and the project consortium, and the Commission cannot be held responsible for any use, which may be made of the information contained therein.

REFERENCES

- Andrighetto, G.; R. Conte; P. Turrini; and M. Paolucci. 2007. *Emergence In the Loop: Simulating the two way dynamics of norm innovation*. Schloss Dagstuhl, Germany, G. Boella, L. v. d. Torre and H. Verhagen (Eds.).
- Andrighetto, G.; D. Villatoro; and R. Conte. 2010. “Norm internalization in artificial societies”. In *AI Communications*, Volume 23 Issue 4, 325-339.
- Andrighetto G.; J. Brandts; R. Conte; J. Sabater; H. Solaz; and D. Villatoro. 2013. “Punish and Voice: Punishment Enhances Cooperation when Combined with Norm-Signalling”. In *PLoS One*, vol. 8 (6), 1-8.
- Arlacchi, P. 1992. *Mafia von Innen: Das Leben des Don Antonio Calderone*. Fischer, Frankfurt am Main.
- Axelrod, R. 1986. “An evolutionary approach to norms”. In *American Political Science Review* 80 (4), 1095-1111.
- Bicchieri, C. 2006. *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press, New York.
- Conte, R. and C. Castelfranchi. 1995. “Understanding the effects of norms in social groups through simulation”. In *Artificial societies: the computer simulation of social life*, G.N. Gilbert and R. Conte (Eds.). UCL Press, London.
- Conte, R.; G. Andrighetto; and M. Campenni. 2014. *Minding norms: mechanisms and dynamics of social order in agent societies*. Oxford university press, Oxford.
- Corbin, J. and A. Strauss. 2008. *Basics of qualitative research, 3rd ed.* Sage, London.
- Dickie, J. 2004. *Cosanostra: A history of the sicilian mafia*. Palgrave Macmillan.
- Hechter, M. and K. D. Opp. 2001. *Social Norms*. Russell Sage Foundation, New York.
- Horne, C. 2007. “Explaining norm enforcement”. In *Rationality and Society* 19(2), 139-170.
- Lotzmann, U.; M. Möhring; and K. G. Troitzsch. 2013. “Simulating the emergence of norms in different scenarios”. In *Artificial Intelligence and Law* 21(1), 109-138.
- Neumann, M. 2008. “Homo socionicus. A case study of simulation models of norms”. In *Journal of Artificial Societies and Social Simulation* 11(4), <http://jasss.soc.surrey.ac.uk/11/4/6.html>
- Savarimuthu, B.; S. Cranefield; M. A. Purvis; and M. K. Purvis. 2013. “Identifying prohibition norms in artificial societies”. In *Artificial intelligence and law* 21(1), 1-46.
- Scherer, S.; M. Wimmer; and S. Markisic. 2013. „Bridging narrative scenario texts and formal policy modelling through conceptual policy modelling”. In *Artificial intelligence and law* 21(4), 455-484.
- Ullman-Margalit, E. 1978. *The emergence of norms*. Oxford university press, Oxford.

Policy Modelling

INTEGRATING OPTIMISATION AND AGENT-BASED MODELLING

Peter George Johnson

Tina Balke

Centre for Research in Social Simulation, University of Surrey, Guildford, GU2 7XH, UK

Lars Kotthoff

Cork Constraint Computation Centre, University College Cork, Cork, Ireland

Email: peter.johnson@surrey.ac.uk

KEYWORDS

agent-based modelling, optimisation, incentive design, photovoltaic, policy modelling

ABSTRACT

A key strength of agent-based modelling is the ability to explore the upward-causation of micro-dynamics on the macro-level behaviour of a system. However, in policy contexts, it is also important to be able to represent downward-causation from the macro and meso-levels to the micro, and to represent decision-making at the macro level (i.e., by governments) in a sensible way. Though we cannot model political processes easily, we can try to optimise decisions given certain stated goals (e.g., minimum cost, or maximum impact). Optimisation offers one potential method to model macro-level decisions in this way. This paper presents the implementation of an integration of optimisation with agent-based modelling for the example of an auction scenario of government support for the installation of photovoltaic solar panels by households. Auction type scenarios of this kind, in which large groups of individuals or organisations make bids for subsidies or contracts from government, are common in many policy domains.

INTRODUCTION

Agent-based modelling (ABM) is an increasingly popular technique in the social sciences to evaluate the effect of policies and other instruments that affect groups of people. This is a result of the fact that ABM is well suited to exploring the macro-level behaviour of a system resulting from the micro-dynamics of the agents represented. We term the aggregate system level or actors that have far and wide-reaching effect “macro” and individual-level, i.e., household, actors and dynamics “micro”. Everything in between is called “meso”. Many heterogeneous and autonomous agents and their interaction with one another, as well as the environment they act in, are relatively easily represented in ABMs and the resulting models represent upward-causation well. However, in policy contexts it is common for downward-causation to also play a central role, and thus it is important for this to be in-

cluded in any policy relevant ABM. Thus, a policy ABM may need to include a ‘policymaker’ type agent, and a process through which its decisions affect other agents. Representing the decision-making process of such a policymaker agent is likely to be difficult, reflecting the political and complex nature of policy-making. One option is to focus on a small selection of central goals a policymaker may have, such as minimising the cost of a policy, maximising the effect, or optimising with respect to another indicator (e.g., environmental quality).

Optimisation technologies are well-established in computer science and artificial intelligence to select the optimal element(s) or solution(s) (with regard to some goals) from some set of available alternatives. In many policy examples, there is a choice of possible decisions to make and actions to take given the current state of the system. Typically, these decision and action alternatives can be distinguished by the “value” they add to the overall goals of the policy makers. In such cases, it is desirable to make the decision that optimises the value with respect to the goal. This is an optimisation problem.

A typical example of an optimisation problem in this context is where a set of companies are putting in bids for contracts to provide a certain service that should be realised as a result of a policy decision. Usually, cost is the primary optimisation criterion – the service should be provided as cheaply as possible. However, the adequacy of the service has to be ensured as well. If only one company can get the contract, the problem is usually easy to solve – the lowest bidder that maintains the required standards gets the contract.

However, in many scenarios, it is not that simple. The service to provide may be complex and consist of several sub-services. Companies bid to provide those sub-services and it is the task of the policy maker to ensure that everything necessary is provided in the end. Instead of simply minimizing the cost, other considerations need to be taken into account now, making the problem much harder to solve. This is where optimisation technology is useful and can make life much easier for the policy maker.

This paper presents a method for the integration of this type of optimisation with an ABM built in NetLogo. This

method offers policy modellers a desirable option for representing decision-making at the macro-level. To illustrate our approach the example case of government support for household installation of photovoltaic (PV) solar panels is used. In this case, households/agents make bids for government support, and the optimisation selects a subset of these and ‘gives’ out support in such a way that the goal set out by the user (e.g. minimization of costs) is achieved. This type of auction scenario is particularly common in many policy domains.

Although it might be possible to implement optimisation routines in NetLogo by embedding them in an agent’s behaviour rules, it should be noted that this would prove highly computationally demanding, thus reducing the model speed considerably. More fundamentally, it would be time inefficient owing to the extensive coding required. It is thus much more efficient to use a stand-alone optimiser and integrate the two.

The paper is structured as follows. In the next section we introduce the individual components of our integrated approach, namely ABM and optimisation. Next, we describe previous work and outline the novelty of our approach against it. Then, the case study of PV adoption in the Emilia Romagna region is presented alongside details of the ABM. In the next section we focus on the main contribution of this paper by describing and discussing the proposed integration approach itself. Finally, some initial results of the final integrated model are presented. The paper closes with a brief conclusion section.

METHODOLOGICAL BACKGROUND

Agent-based modelling

Squazzoni, Jager, and Edmonds (2013); Chattoe-Brown (2013); Gilbert (2008) provide overviews of ABM and their use in the social sciences. ABM is a form of simulation modelling in which multiple agents act and interact within an environment. Agents can represent any decision-making unit (e.g., person, household, firm) and are autonomous, can communicate with each other, and are typically heterogeneous. The agents act and interact within an environment which may represent an abstract conceptual space (e.g., social or opinion space), a real physical or geographic space (e.g., a building or country), or have no real meaning (i.e., when agents are in a network connected by links). Typically the agent behaviour rules, attributes and the environment are setup using empirical data or theory, and the model is then run. The emergent macro-level patterns that are the output of the model are then compared and analysed alongside the micro-level rules. Broadly, ABMs can be used first, to explore and explain the mechanism of a theory of individual behaviour on the whole system, or second, to describe and forecast the behaviour of a system, or third, in a participatory context to explore a system and its behaviours with stakeholders. ABM is most often used when the researchers: (i) are interested in modelling interactions and feedback between actors, and actors and their environment, (ii) believe heterogeneity of actors is important

in the system, (iii) are interested in the spatial dynamics of a system, (iv) believe path dependence may be an important element in the system, (v) believe actors in the system have behaviours that change, or adapt over time, and/or (vi) want to use an intuitive and flexible modelling approach for participatory modelling (Johnson, 2014).

Whilst ABM are well suited, and typically used, to represent the decision-making and behaviour of many heterogeneous micro-level agents, it is less common to represent meso and macro-level actors. At the meso-level, examples might include firms, or government agencies, whereas at the macro-level examples could include governments or nations. However, it is difficult to endow these agents with sophisticated behaviours (beyond profit/utility maximisation or heuristics), especially when interacting with a large number of micro-level agents. Indeed, if the behaviour is more complex, or the number of micro-level agents is large, the process is likely to be computationally demanding. This reflects the problem that the meso or macro-level agent has to perceive, use and manipulate information from all of the agents in the simulation.

The combination of the facts that meso and macro-level behaviour is worthy of representing, and that it can prove difficult to implement in NetLogo means further development is an important avenue for the continued evolution of ABM using NetLogo for policy and social science applications.

Optimisation

Optimisation aims to find the solution to a problem that is optimal with respect to an application-specific criterion. It is applicable in a wide range of contexts and a well-known technique in many sciences. Examples include minimising waste in a production pipeline that can manufacture a range of different items from the same raw material for a set of orders, minimising the cost of travel while visiting a list of locations, or maximising the value while minimising expenditure when bidding on a list of items with associated cost and value.

In a policy-making context, optimisation problems can arise in a number of scenarios. The budget allocated to achieve a certain objective may need to be distributed across different policy instruments in an optimal fashion. The provision of a service may rely on several companies putting in bids to provide this service for a certain fee. A regional development project may want to minimise the negative impact on the environment.

Optimisation is a large field in its own right, and a survey of the different areas is beyond the scope of this paper. The interested reader is referred to Chong and Zak (2008) for an introduction. There are a number of mature software packages that implement optimisation technologies that can be used here.

It is important to remember that optimisation facilitates decision-making at the macro-level. Once, for example, the bids for a service are known, the decision based on them can be optimised. To get these bids, other techniques are required – in our case, agent-based modelling.

Without this separate component, optimisation cannot do its work as it would have no data to base its decisions on. It is the integration of agent-based modelling and optimisation that can support the policy maker.

PREVIOUS WORK

There is a relatively limited literature on the integration of optimisation and ABM in the way presented here (i.e., optimisation used in a macro-level/policymaker agent). Broadly, there are three most common forms of integration of optimisation and ABM in previous works: (i) optimisation used as a calibration and validation tool for ABMs, (ii) ABMs used to solve optimisation problems, and (iii) optimisation used in economic ABMs to represent constrained maximisation.

In the first form, which appears the most popular (Franklin, 2012), Terano and Naitoh (2004) use optimisation to fine tune agent parameters in an ABM of marketing strategies. Rogers (2004) and Gilli and Winker (2003) again use optimisation in a similar way in their respective models of financial markets. Calvez and Hutzler (2007) propose using ‘adaptive dichotomic optimisation’, which represents a form of single objective genetic algorithm optimisation, and demonstrate with an ABM of a financial market. Narzisi, Mysore, and Mishra (2006) use optimisation in the same way to calibrate micro-level parameters, but also to help optimise the emergent behaviour in their model of disaster management (i.e., to minimise casualties, and other indicators), and in effect to explore the results of their model. Schutte (2010) uses optimisation to calibrate their model of air traffic management and suggests the approach be used elsewhere. Jakob et al. (2012) use optimisation in a related form, in which an ABM and optimisation model of anti-pirating techniques for shipping companies are developed and used to validate and complement each other within a larger tool.

In the second form, Barbati, Bruno, and Genovese (2012) present a review of the uses of ABMs, and broadly agent approaches, to help solve optimisation problems. They identify two types of agents in this sense, first physical agents that may represent physical entities such as workers or machines, and second, ‘functional’ agents, that represent nothing in the physical world, but are a piece of software used to carry out subtasks of the ABM. They also identify two structures: one in which many agents self-organise to solve a problem, or a second in which there are ‘mediator’ type agents which set an optimised plan, that may be refined by the ‘worker’ agents. This type of integration is less formal than the first and is essentially the application of ABM to scheduling or resource allocation problems. Examples include Weichhart et al. (2002), Davidsson, Persson, and Holmgren (2007), and Socha and Kisiel-Dorohinicki (2002).

Davidsson, Persson, and Holmgren (2007) present the embedding of optimisation in micro-level agents to aid resource allocation problems, using an example taken from the food industry. This is one of the more similar integrations to that which is described in this paper, however there are some crucial differences. First and most im-

portantly, the actors represented are at the micro-level, rather than at the macro-level (i.e., policymakers). Second, the intention of the model is to solve some production / resource problem, rather than to represent decision-making accurately. Third, the implementation is done using one piece of software, rather than integrating an existing ABM with optimisation software.

In the third form, there is a stream of work from the field of ACE (agent-based computational economics), that uses optimisation to represent standard utility maximisation, with constraints. Chakraborti et al. (2011) and Chen, Chang, and Du (2012) provide recent reviews of this work. Here, the aim is to represent the economic theory of utility, or profit, maximisation, often of macro-level agents (e.g., central banks, government). The agents typically face constraints on their maximisation, and so this becomes an optimisation problem. Again, this is one of the more similar streams of work to that presented here. However, the ABM here is not solely an economic model, the agent behaviours include social and policy factors.

A similar type of optimisation within agents has also been used in ABMs to explore social dilemma questions (Gotts, Polhill, and Law 2003 provide a review of some of this work), in which the approach is used to represent rationality in the sense of utility maximisation.

What separates these previous works and this paper is that the approaches use either, (i) optimisation to analyse the results of the simulation after it has been run, (ii) embed the optimisation in agents architecture to represent economic decision making, or (iii) use agent type approaches on common optimisation problems. The approach described in this paper allows a run-time integration of optimisation and ABM, where the optimisation component (representing a macro-level policy agent) can communicate with, and influence, and thereby optimise macro-level decisions in the simulation during its execution.

CASE STUDY

The example used in this paper is that of the ‘ePolicy social simulator’ developed as part of the ePolicy FP7 project on engineering the policy life-cycle (see <http://www.epolicy-project.eu/> for more details) which uses the Emilia Romagna region in Italy as its case study. Italy has few fossil fuel resources and relies heavily on imported natural gas, which is why (together with a general sentiment of the Italian population against nuclear power) renewable energy has long been a topic of interest in Italy. Of the different renewable energy technologies available, PV panels have been of particular interest in Italy due to climate and economic conditions that have resulted in a steep rise in capacity in Italy in the last couple of years. The regional government in Emilia Romagna is particularly interested in the potential for new technologies to contribute to energy production. This interest in PV serves as the basis of the use of Emilia Romagna as the case study.

The ePolicy social simulator, which is an ABM, has been developed to serve as a component of a wider de-

cision support system (DSS) for policy makers and analysts working on energy policy. It is intended to be used to answer the policy question, “What are the effects of different policy instruments on PV system diffusion in the Emilia Romagna region?”. For Emilia Romagna, two specific regional policy instruments have been identified in collaboration with regional policy-makers: investment grants and interest-rate support (for interest on loans to purchase a system). The ABM simulates the behaviour of households in reaction to these policy support schemes. The agents’ behaviour rule simulates the consideration of, and decision to install, PV and is based upon a household survey and interview data from Italy. When agents decide to install PV, they may make a bid to the regional government to apply for support either in the form of grants or interest-rate support. This form of application is motivated by similar schemes that the Emilia Romagna region has run in the past. The optimiser then has to decide which bids to fund, given the overall budget and power capacity target and any other goals. The synthetic population of agents in the ABM is setup using population data from Italy. The environment of the ABM in which the agents interact in is setup using GIS data of the Emilia Romagna region. The outputs of the ABM are the aggregate costs of installations, aggregate power generated by PV, and total number of installations.

INTEGRATION APPROACH

The integrated approach is implemented in a Java-based user interface which unifies and encapsulates both the ABM and the optimiser. Supplementary material in form of a more detailed description of the implementation, experimental results, and where the software can be downloaded be found at <http://www.epolicy-project.eu/sites/default/files/public/D5.3.pdf>. While NetLogo is used for the ABM, the Ipsolve software (see <http://lpsolve.sourceforge.net/> for details) provides the optimiser. Both provide application programming interfaces (APIs) that we utilise in the user interface.

The purpose of the user interface is to transparently make use of the relevant technologies without burdening the user with additional parameters and setup steps. Indeed, there is no indication of how the specified problem is being solved underneath. This also allows for easy and transparent integration of alternative solving methods.

Figure 1 describes the integration of the ABM and optimisation. Using the user interface, the user first defines the scenario they wish to explore. The scenario options include selecting the region (Emilia Romagna or just its capital, Bologna), budget and its distribution over time (first come, first served, even, ramp up or down), intended PV power supply, optimisation criterion (minimising the budget spent, maximising the power production, maximising the participation, i.e. the number of funded bids), and their beliefs about the status of national level policy instruments (feed-in tariffs and income tax liability reductions). The simulator and its data are then loaded, and the optimiser is informed of the user’s choices on optimisa-

tion criterion, budget and budget distribution.

Now the simulation begins. For each time period (one year), household decisions are made, and a list of households who wish to bid for regional government support is generated along with the bid amount and the size of the PV installation. This list of bids is then passed to the optimiser which builds an optimisation model based on the bids and the parameters specified by the user.

The optimisation model is a variant of the so-called knapsack problem – given a set of items, each with a certain weight and value, and a knapsack with a certain capacity, maximise the value of the items put in the knapsack without exceeding its capacity. In the case of maximizing the power produced for example, the weight of the items is the cost, the value the power produced, and the capacity of the knapsack is the budget.

Knapsack-type problems are very common in the optimisation domain and can be solved very efficiently in practice even though they are hard to solve in general. In our experiments, the optimisation problems were solvable within a fraction of a second for all scenarios and optimisation criteria.

The optimiser solves the optimisation problem and generates a list of funded bids. This funded list is then passed back to the ABM where the households implement the decisions and install PV, resulting in costs to the policy instrument, and increased power generation. This process is repeated until the simulation has reached the desired final year (currently 2020). Finally, the produced power, spent budget, count of funded bids (i.e., installations) is provided to the user.

INITIAL RESULTS

Table I presents a comparison of the key outputs that the different optimisation objectives and methods achieve on the data of an auction the Emilia-Romagna region ran in 2004, with a budget of 3,200,000 Euro and a target power of 1,200 kW. All scenarios use the first come, first served budget distribution and have both grant and interest payment incentives enabled. The values are the ones achieved after the first simulation step.

The ordering approach corresponds to what the Emilia-Romagna region did in the 2004 auction to determine the winning bids. All bids are ranked according to criteria that depend on the optimisation function. Then the list of bids is traversed in this order, with each bid funded if there are sufficient funds available and the power target has not been reached. For example, when maximising the power production, the bids are ranked by the ratio of power production of the installation divided by its cost.

We are showing only the results after the first time step of the simulation because the results across all time steps are not comparable. The optimisation takes into account all the information available at each step and makes the optimal decision for that. This is what we aim to demonstrate here. Ideally, we would be able to make decisions based on the information for the entire simulation – while funding a particular bid at time step n may appear to be a good decision, it is possible that at time step $n + 1$ a better

bid will be made that should be funded instead. However, it is impossible for us to do this – decisions have to be made for each time step because earlier decisions will affect the course of the simulation and later time steps.

The approaches that do not use optimisation tend to leave more “slack” at each time step, i.e. not all the bids that could be funded are. This means that more of the budget is available in subsequent time steps, allowing to fund potentially better bids. It is because of this that the approaches that do not use optimisation may achieve overall better results than the optimisation approach. This is, however, purely by chance – optimisation is overall likely to obtain better results, especially when the number of time steps is large and decisions are spread out across them. Here, we compare the numbers after the first simulation step to remove this random element.

Table I clearly shows that using optimisation technology is worthwhile. The simple first come, first serve approach for deciding which bids are funded achieves significantly less power production and fewer funded bids than the other approaches. Ranking the bids depending on the optimisation criterion achieves much better results. Yet, using optimisation is able to improve even further on this. These results clearly demonstrate the benefit of using optimisation technology for ABM.

Note that in the case of optimising to maximize participation, the order and the optimisation approach achieve the same number of funded bids, but the order approach at a lower cost. This is because the optimisation approach considers only the single objective of maximizing the participation – there are several assignments of funds to bids that achieve the same participation and the optimisation approach happened to choose the one with a higher cost.

This is not an inherent limitation of the optimisation approach – to take the cost into account as well, we can model the problem as a so-called multi-objective optimisation problem. This class of problems is similarly well studied in the literature and can be solved efficiently in practice. We do not consider this approach here simply for the sake of simplicity of the exposition – multi-objective optimisation problems are by their very nature more complex to define.

CONCLUSION

We have made a first step towards integrating agent-based modelling and optimisation technologies at runtime. In many scenarios, a special type of agent is required to make downward-causation decisions that affect the other agents in a simulation. Such decisions need to take into account the current state of the simulation and optimise for a criterion. Optimisation technology is a natural choice for implementing this process. Instead of integrating the optimisation as part of the decision-making of a ‘governmental agent’ within the ABM, our prototype linked the ABM and an existing solver for optimisation questions and allowed them to communicate during the course of the simulation.

In our case study for funding photovoltaic installations in Italy’s Emilia-Romagna region, our integrated simula-

tion has shown significant improvements on previous results by using the optimisation component instead of less sophisticated approaches.

Although we present the integrated ABM-optimisation approach with the help of the Emilia Romagna PV adoption case study in this paper, our approach is not limited to this case study and can be applied to a large variety of topics. It is in particular useful when macro-level decision-making and its influence on the agent decisions are of importance for the simulation. With our approach one can both use the ABM to understand and analyse the agent decision-making behaviour at the micro-level and at the same time to make optimisation decisions at the macro-level. As noted before, because of the interaction of the two components during the execution of the simulation, the macro-level decisions influence the agents at the micro-level and therefore influence the whole simulation result. This interaction between the different levels allows for the study of the complete system, as well as the interaction between levels.

Our future work can be divided into three streams. The first stream is the extension of the ABM and the exploration of other scenarios for the Emilia Romagna case study. In particular, applying the same methodology to an entire country, such as Italy, would be of practical interest. The second stream concerns the generalization of the integration methodology and its application to other domains, as well as refining the current interactions between the components. In particular, multi-objective optimisation problems and the effect of optimising multiple, independent decisions at each time step on the entire simulation can be explored. Thirdly, exploring the implementation potential of the optimisation component alone may be worthwhile pursuing. This would involve applying the component to real world data on bids for government support, under a range of case studies, and engaging with potential policy users to understand their objectives and constraints.

ACKNOWLEDGEMENTS

This research was funded under the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 288147 (ePolicy).

REFERENCES

- Barbati, Maria, G. Bruno, and A. Genovese (Apr. 2012). “Applications of agent-based models for optimization problems: A literature review”. In: *Expert Systems with Applications* 39.5, pp. 6020–6028. ISSN: 09574174. DOI: 10.1016/j.eswa.2011.12.015. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0957417411016861>.
- Calvez, Benoît and Guillaume Hutzler (2007). “Adaptive Dichotomic Optimization: a New Method for the Calibration of Agent-Based Models”. In: *European Simulation and Modelling Conference (ESM 2007)*, pp. 415–419.

- Chakraborti, Anirban et al. (July 2011). "Econophysics review: II. Agent-based models". In: *Quantitative Finance* 11.7, pp. 1013–1041. ISSN: 1469-7688. DOI: 10.1080/14697688.2010.539249. URL: <http://www.tandfonline.com/doi/abs/10.1080/14697688.2010.539249>.
- Chattoe-Brown, Edmund (Aug. 2013). "Why Sociology Should Use Agent Based Modelling". en. In: *Sociological Research Online* 18.3. URL: <http://www.socresonline.org.uk/18/3/3.html>.
- Chen, Shu-Heng, Chia-Ling Chang, and Ye-Rong Du (Apr. 2012). "Agent-based economic models and econometrics". English. In: *The Knowledge Engineering Review* 27.02, pp. 187–219. ISSN: 0269-8889. DOI: 10.1017/S0269888912000136. URL: <http://journals.cambridge.org/abstract/S0269888912000136>.
- Chong, E. K. P and S. H. Zak (2008). *An Introduction to Optimization*. 3rd. Wiley-Interscience.
- Davidsson, Paul, Jan A Persson, and Johan Holmgren (2007). "On the Integration of Agent-Based and Mathematical Optimization Techniques". In: *Agents and Multi-agent systems: Technologies and Applications*.
- Franklin, Chris (2012). "Multi-objective Optimisation using Agent-based Modelling". PhD thesis. Stellenbosch University.
- Gilbert, Nigel G. (2008). *Agent-based Models*. Sage Publications Ltd., p. 112.
- Gilli, M. and P. Winker (Mar. 2003). "A global optimization heuristic for estimating agent based models". In: *Computational Statistics & Data Analysis* 42.3, pp. 299–312. ISSN: 01679473. DOI: 10.1016/S0167-9473(02)00214-1. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0167947302002141>.
- Gotts, N M, J G Polhill, and A N R Law (2003). "Agent-Based Simulation in the Study of Social Dilemmas". In: *Artificial Intelligence Review* 19, pp. 3–92.
- Jakob, Michal et al. (2012). "Agents vs . Pirates : Multi-Agent Simulation and Optimization to Fight Maritime Piracy". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Innovative Applications Track (AAMAS 2012)*.
- Johnson, Peter George (2014). "The SWAP Model: Policy and Theory applications for Agent-Based Modelling of Soil and Water Conservation Adoption". PhD thesis. University of Surrey.
- Narzisi, Giuseppe, Venkatesh Mysore, and Bud Mishra (2006). "Multi-Objective Evolutionary Optimization of Agent-Based Models: An Application to Emergency Response Planning". In: *Proceedings of Computational Intelligence*.
- Rogers, Alex (2004). "Multi-objective Calibration for Agent Based Models". In: *Proceedings of 5th Workshop on Agent-Based Simulation*. ISBN: 3936150311.
- Schutte, Sebastian (Jan. 2010). *Optimization and Falsification in Empirical Agent-Based Models*. en. URL: <http://jasss.soc.surrey.ac.uk/13/1/2.html>.
- Socha, Krzysztof and Kisiel-Dorohinicki (2002). "Agent-based evolutionary multiobjective optimisation". In: *Proceedings of the 2002 Congress on Evolutionary Computation*. Vol. 1. Ieee, pp. 109–114. ISBN: 0-7803-7282-4. DOI: 10.1109/CEC.2002.1006218. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1006218>.
- Squazzoni, Flaminio, Wander Jager, and Bruce Edmonds (Dec. 2013). "Social Simulation in the Social Sciences: A Brief Overview". In: *Social Science Computer Review*. ISSN: 0894-4393. DOI: 10.1177/0894439313512975. URL: <http://ssc.sagepub.com/cgi/doi/10.1177/0894439313512975>.
- Terano, T. and K. Naitoh (2004). "Agent-based modeling for competing firms: from balanced-scorecards to multiobjective strategies". In: *Proceedings the 37th Annual Hawaii International Conference on System Sciences*. IEEE. ISBN: 0-7695-2056-1. DOI: 10.1109/HICSS.2004.1265251. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1265251>.
- Weichhart, Georg et al. (2002). "Modelling of an Agent-Based Schedule OPTimization System". In: *Proceedings of the Fourth Congress on Evolutionary Computation*. Penya 2002.

AUTHOR BIOGRAPHIES

PETER GEORGE JOHNSON is a Research Fellow at the Centre for Research in Social Simulation. His interests revolve around the use of agent-based modelling to address environment and energy-related policy issues. He is also interested in the use of models in policy more broadly.



TINA BALKE is a Postdoctoral Research Fellow at the Centre for Research in Social Simulation. Her research focusses on modelling the impact of norms, different institutional settings and policy instruments on (human) decision making in complex settings and systems. In her research she combines empirical social science methods, agent-based models, as well as logic-based approaches and artificial intelligence techniques.



LARS KOTTHOFF is a postdoctoral researcher at University College Cork, Ireland. He took up a position there after finishing his PhD at the University of St Andrews, Scotland. His research interests are in combinatorial optimisation, machine learning, and how to combine the two fields. He is involved in several multi-disciplinary efforts that aim to bridge the gap between different disciplines and exploit synergies through cross-fertilisation.



TOWARDS AN AGENT-BASED MODEL ON CO-DIFFUSION OF TECHNOLOGY AND BEHAVIOR: A REVIEW

Thorben Jensen
Future Energy and Mobility Structures
Wuppertal Institute
Doeppersberg 19, 42103 Wuppertal, Germany
E-mail: Thorben.Jensen@wupperinst.org

Emile Chappin
Energy and Industry, Faculty TBM
Delft University of Technology
Jaffalaan 5, 2628 BX Delft, The Netherlands
E-mail: E.J.L.Chappin@tudelft.nl

KEYWORDS

Energy efficiency, behavior, agent-based modeling, social simulation, Transformational Products.

ABSTRACT

In this paper we propose an agent-based modeling study to assess policy options towards technological interventions to energy consumption behavior, in particular products also known as ‘Transformational Products’. They are novel domestic heating energy efficiency enabling devices, designed to change heating behavior of their users. Innovative behavior can then trigger social learning, i.e. propagation of changed behavior from users to their peers. Arguing that effects of such domestic technology needs to be assessed beyond the household scale, we state the following requirements for a simulation model: it should be an agent-based model considering co-diffusion of behavior and technology. Bottom-up, it should base upon households of certain lifestyles, which are connected in social networks. Model rules should be rooted in socio-psychological theory. Model validation should be feasible. We conduct a literature review on modeling studies, evaluating to which extent these meet these stated requirements, and conclude: (1) an agent-based modeling study meeting all these requirements is not found, suggesting that it does not exist yet; (2) a combination of existing models would meet the stated requirements. A use case for the proposed simulation model for policy makers is enabling strategic decisions by comparing the effects of technology that addresses heating behavior with technology which automatizes heating.

INTRODUCTION

To combat climate change and to deal with the depletion of fossil resources, increased energy-efficiency in heating is urgently needed. Domestic heating may play a key role in the energy transition, as it provides for roughly 28% of energy consumption in the EU-15 (Balaras et al. 2007). Of various possible approaches to decarbonize heating, change in domestic heating behavior is appealing due to its low expected investment costs, its resource-efficiency and broad applicability in the housing sector. The potential for energy savings of households by behavior change of ca. 30% is significant (Peschiera et al. 2010). Additionally, change in heating behavior additionally bears the potential to spread via

social learning. In this paper we are exploring ways for assessing the effects of efficiency enabling devices more holistically.

Effects of technological interventions to domestic heating behavior are commonly investigated on the scale of households, which leaves spatial up-scaling to a larger area as an open challenge (Ernst 2014). However, crucial factors to the intervention potential lie beyond the household level: (1) the technology diffusion process (e.g. adoption rates, product market shares, ideally specified for societal groups) between consumers and (2) behavior diffusion, (e.g. innovative behavior stimulated by such devices) which can spread in social networks via social learning. We coin the combination as *co-diffusion of technology and behavior* in socio-technical systems. This emergence of diffusion suggests that investigating efficiency appliances on the household scale is not sufficient. Rather, their impact should be assessed dynamically and on a societal scale.

To realize the energy saving potential of heating efficiency enabling devices, policy options for their diffusion need to be evaluated. Existing knowledge regarding policy impact on diffusion (see Tao Zhang and Nuttall 2011) may be transferred to the co-diffusion of efficiency products and the behavior changed by these. We need to understand co-diffusion of behavior and technology, i.e. heating efficiency enabling devices, and the opportunities for policy to manage that diffusion. In this paper we aim to find out if and which simulation models are suited for this task. Therefore, we review relevant modeling studies.

The structure of this paper is as follows. First, the potential role of heating efficiency enabling devices is described. From this, requirements for a simulation model assessing their potential are derived. Second, a literature search for simulation models meeting these requirements is conducted, starting separately with models on diffusion of energy efficient behavior and efficiency technology. Thereafter, underlying assumptions of the key publications in this realm are examined. Finally, implications for policy makers and stakeholders are given, followed by conclusions and an outlook on further research.

THE POTENTIAL ROLE OF HEATING ENERGY EFFICIENCY ENABELING DEVICES

Here, established paradigms of heating efficiency enabling devices are presented, using Transformational Products as a case technology. Finally, requirements for a simulation model assessing their possible effects are derived.

Behavior theory widely accepts the fact that routinized behavior (or ‘bad habits’) is a key barrier to more sustainable heating practices (Jackson 2005; Jaeger 2003). Transformational Products particularly address ‘bad habits’ by causing *friction* to create *situative awareness*, which enables the user to make choices where, otherwise, routinized behavior would take place. As a comprehensive example, Laschke et al. (2011) present the ‘never hungry caterpillar’, a caterpillar-like device that is supposed to be placed next to a TV. If the TV is switched to stand-by, it twists and thus symbolizes discomfort, which creates awareness on the waste of energy. Thus, awareness is created just in time and can immediately be translated into action. Enabling choice can further be combined with ‘nudging’ (Thaler and Sunstein 2009) towards more sustainable behavior. As soon as the TV is turned off completely, the caterpillar stops and thus stops showing unease. Like this, Transformational Products have the potential to support behavior change by *unfreezing* routines (see Lewin 1951) and to close the gap between consumers’ intentions and actions. Afterwards, refreezing new routines is possible, given a sufficient frequency of the new action and strength of reinforcement (Jager 2003). Figure 1 illustrates the functioning of Transformational Products, including the potential of behavior diffusion.

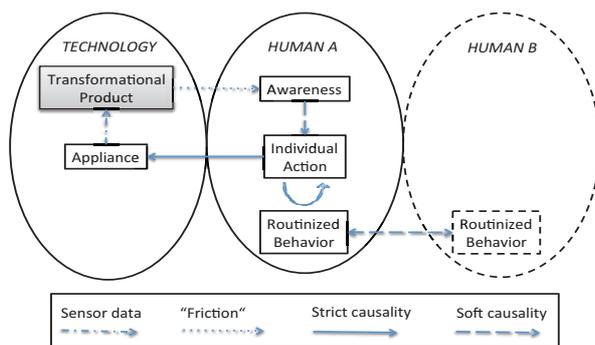


Figure 1: Effects of Transformational Products (partly based on Wood and Newborough 2003, fig. 4)

As the concept of Transformational Products is novel, prototype testing intervening in heating routines is just starting. Results on their energy saving potential are expected for the first quarter in 2015. Again, because Transformational Products address heating behavior, energy savings for adopters can yet be estimated as up to 30% (see Peschiera, Taylor, and Siegel 2010), as for similar interventions.

Difference from other efficiency enabling devices

Transformational Products are designed to solve the shortcomings of the technological paradigms *automation* and *persuasion by information*.

Automation is aimed at ruling out behavioral deficits and making users obsolete, e.g. with programmable thermostats. A drawback is that consumer behavior often interferes with automation systems, e.g. by opening windows while heating. Also rebound effects are a common consequence of automation systems, e.g. programmable thermostats can cause longer heating times and consequently higher energy consumption, compared to manual control (Guerra-Santin and Itard 2010). Finally, automation only works where installed, e.g. heating at home do not benefit from automated heating at work.

Persuasion by information emphasizes the pivotal role of users, intending to change their behavior by presentation of information, e.g. via a Smart Meter display. This can include normative comparison to peer consumption (see *Watt's Watts* device at Gulbinas et al. 2014). Persuasion uses the successive steps *monitoring*, *visualization* and *user awareness* (Laschke et al. 2011). Unfortunately, awareness alone does not effectively translate into behavior change – particularly not on the long run (Wood and Newborough 2003). Presumably, this is caused by persistence of routines. Thus, if not addressed repetitively and with the right timing, learning of different heating routines might be of low success (Jager 2003).

Transformational Products attempt to address the drawbacks of both. First, they address consumer behavior instead of eliminating it. Second, they aim at changing behavior routines situatively. Behavior change is therefore likely to be quicker and more persistent. Persistently changed heating routines could thus be more transferable, e.g. beyond the time of product usage and between locations.

Requirements for Agent-based Modeling of Transformational Products

Because Transformational Products are currently in the prototyping phase, their performance is only observable for small sample sizes of households. Simulation modeling would enable exploring their possible effects already before market launch. Perfectly predicting ‘the future’ is not the primary focus, because of high uncertainty of future product diffusion at market initialization (Rogers 2010), uncertain future consumer preferences and the complexity of socio-technical interactions. Instead of predicting, a simulation model can be useful for exploring possible futures.

For this purpose, a good choice for exploring emergent system behavior resulting from complex socio-technical interactions is agent-based modeling (Chappin 2011). This is because processes of product diffusion

among cognitive agents cannot be solved analytically (Schwarz 2007), and diffusion of technology and changed behavior is an emergent process, which is naturally analyzed bottom-up, based upon the final units of empirical assessment, i.e. in this case households. Furthermore, when giving policy advice with a simulation model, this should consider human behavior (Sopha et al. 2011). To this, an advantage of social simulation to estimate policy effects to technology diffusion is to gain valuable time for decision making. Agent-based modeling in particular has the advantage to explicitly model agents' heterogeneity (Grimm et al. 2006), which can be used to reveal heterogeneous effects of a technology throughout society.

Consequently, we propose analyzing the diffusion of Transformational Products and heating behavior with an agent-based model. The purpose of an according study should be to explore potential future alterations of domestic heating behavior by this technology. Such a simulation model should allow assessing policies supporting the diffusion of Transformational Products. This would enable decision makers to decide on policy as soon as possible. The suggested simulation model should meet the following design requirements: it should be (1) an agent-based model, (2) households should be its focal units of analysis, i.e. its agents, including (3) lifestyles to address the 'up-scaling problem' (see Ernst 2014). (4) It should consider social networks. (5) Assumptions on model rules should explicitly base on acceptable socio psychological theory. (6) Validation should be addressed with respect to the applicability of such models for policy makers to decide on policies to enable Transformational Products to flourish and be effective. Such an analysis would benefit if (7) co-diffusion of technology and behavior as an important underlying mechanism is included.

EXISTING DIFFUSION MODELS

A literature review is conducted in order to, first, investigate if the here proposed modeling study is novel, i.e. whether an existing agent-based modeling study yet could describe the co-diffusion of both Transformational Products and sustainable heating behavior. The second aim is to guide modeling decisions of future work. Thus, successes and limitations of the reviewed fields are assessed. An additional output of this is identification of suiting approaches for *TAPAS validation*, i.e. "tak(ing) a previous model and add(ing) something" (Frenken 2004; Windrum et al. 2007). With TAPAS validation, existing models can potentially be used as 'ingredients' for building the proposed model.

To find relevant papers, we select publications meeting several of our model requirements. Publications being cited by or citing this set, according to SCOPUS and Google Scholar are first reviewed for closeness to the proposed agent-based model. If this is the case, their undirected citation network is explored, too. The only exception to this is made for reviewing the Consumat

framework (see Jager and Janssen 2012), which we do generally, instead of for all its applications.

Agent-based Modeling Studies

We separately present agent-based modeling studies on diffusion of behavior and technology. Each study is scored to the above stated requirements (see table 1).

Behavior Diffusion.

Zhang et al. (2011) model the diffusion of energy saving awareness. The purpose is to compare the potentials of improving technology and behavior. Agents interact in a 'small-world' network, representing social structure in an office building. However, model rules are not transparently stated. Thorough analysis of the provided model code would be needed to get this information. Model calibration and validation bases on (patterns of) quantitative data. Potential shifts in consumer behavior are not dynamically described over time. Furthermore, individual assumptions are debatable, e.g. that e-mail communication is the main pathway of behavior diffusion.

Azar and Menassa (2014; 2010; 2011) model consumer diffusing practices in order to understand changing energy consumption behavior in groups. At their latest work, they apply the of theory of Relative Agreement by expressing agents' energy consumption by its mean and variance. Including the latter is a simple psychologically-based approach to consider strength of habits: the weaker a consumption habit, the greater its variance, the greater the potential for change. Simulations converge against an equilibrium, i.e. modeled behavior freezes eventually. This conflicts with the empirical finding of the frequent relapse pattern of consumers after interventions (see Peschiera, Taylor, and Siegel 2010). The outcomes are not validated against an empirical case.

Chen et al. (2012) model the effect of *peer-feedback devices* on energy efficient behavior in order to understand the influence of social network characteristics on diffusion of behavior in social groups. The model is empirically-based on the case study by Peschiera et al. (2010). Behavior is modeled as 'energy consumption', underlying a drift towards the consumption level of peers. The empirical 'relapse pattern' of increased energy consumption after an intervention is not reproduced by this model which is, according to the authors, due to insufficient mechanistic understanding. The model assumes complete penetration of *peer-persuasion technology* in the social network, which is a best case scenario. Successive market diffusion of technology would have been more realistic.

Anderson et al. (2014) investigate the influence of types of social networks on diffusion of consumer practices in a generic agent-based model. In simulation experiments, the effect of *environmental champions* (change agents) is assessed. Model validation is only conducted structurally. Like at Azar and Menassa

(2014), simulations converge towards a steady state which could be problematic for validation (see above).

We argue that the modeling methods given at Chen et al. (2012), Anderson et al. (2014) and Azar and Menassa (2014) are each particularly suited to be developed further in a TAPAS approach for the above proposed modeling study. Chen et al. (2012), though validated, tailor their model to peer-persuasion technology. Anderson et al. (2014) and Azar and Menassa (2014) both describe behavior diffusion without specific technological interventions in mind. This could make them applicable to a broader range of efficiency technology, including Transformational Products. The latter captures habits, an important element for describing change to repetitive behavior (Jager 2003).

Technology Diffusion.

Schwarz (2007) and Schwarz et al. (2009) model the diffusion of domestic water saving technology. The social network is based on lifestyle groups, yet with household agents each representing all households of each lifestyle for each km². The modeling choices are based on surveys. The degree of aggregation can be seen as appropriate for yet established product innovations and at intermediate stages of diffusion (see Rogers 2010). For initial diffusion however, a smaller resolution could possibly be more appropriate.

Sopha et al. (2013; 2011) investigate the spread of wood pellet heating systems in Norway. Agents are representing 270 households investigated with a survey. Thus, decisions on model structure, calibration and validation are data driven. For translating empirical findings into the model, three clustered lifestyle groups are used. The study adds weight of evidence to the initial assumption that agent-based models on technology diffusion in social networks can have predictive power. However, adoption of wood pellet stoves could be significantly different from adoption of Transformational Products, thus direct transfer may not be justified.

Kroh et al. (2012) model the diffusion of green electricity. Choice of an electric utility is modeled on household level and is linked to lifestyle groups in a spatial social network. Network construction bases on Holzhauer et al. (2013), which leads to higher connectivity within lifestyle groups. The LARA framework, a psychologically based decision model, is applied (Briegel et al. 2014). Regarding transferability to domestic heating efficiency appliances, several barriers show: (1) utility choice may not be sufficiently similar to heating efficiency appliances; (2) until today, methods of this study are only published to limited extent.

Zhang and Nuttall (2007) model the potential futures of Smart Meter diffusion in the UK, applying the ‘Theory of Planned Behavior’. The focus of their study is to investigate the effect of management strategies and economic regulations on diffusion success. A suggested

agent-based co-diffusion model could at one point be tested for similarity in sensitivity to these policies.

Jager and Janssen (2012) propose the Consumat framework as a meta framework for social simulation. It has been applied in several technology diffusion studies, e.g. for modeling market introduction of green products (Janssen and Jager 2002). The Consumat applications regarding technology diffusion focus on the competition of products (e.g. the diffusion of green products among established grey products). The applicability of this framework seems proven for competing alternative products. It is yet uncertain, however, to which extent it is applicable to novel products like Transformational Products, for which no established market exists. Nevertheless, we acknowledge that the Consumat framework is a highly adaptable meta model. Thus, it could in principle be adapted to model behavior diffusion or to include lifestyles of consumers. However, the downside of this high flexibility is twofold. First, a flexible modeling framework may be too general to a particular problem, which may limit its applicability to co-diffusion of technology and behavior. Furthermore, flexibility comes with a higher number of parameters, which increases empirical work needed for parameterization (see Jager and Janssen 2012). Concluding, the Consumat framework is useful for developing a model on co-diffusion of technology and behavior in the way that it can provide input for the structure of the model, rather than for detailed model parts.

Summarizing, among the reviewed technology diffusion models, the ones presented by Schwarz (2007) and Zhang and Nuttall (2011) appear the most suited to become integrated into a model on co-diffusion of technology and behavior. Others, e.g. the Consumat framework, might serve more indirectly by means of modeling guidance and structural validation.

Applied Diffusion Theories.

In this subsection, we discuss socio psychological theory applied in the reviewed studies. Because modeling co-diffusion of technology and behavior should neither build upon conflicting nor arbitrary grounds, it is important to make sound and explicit modeling decisions.

The larger part of the reviewed studies on technology diffusion use the Theory of Planned Behavior (e.g. N. Schwarz and Ernst 2009; Nina Schwarz 2007; Sopha, Klöckner, and Hertwich 2013; Tao Zhang and Nuttall 2007). This is an acceptable choice for modeling technology diffusion (see Tao Zhang and Nuttall 2011). For modeling behavior diffusion, however, we argue that it is not particularly suited. This is because it frames peer interaction as limited to normative pressure and, implicitly, information (Fishbein and Ajzen 2005). However, persistence of habits (see Jackson 2005; Peschiera, Taylor, and Siegel 2010) is not considered.

The reviewed studies on behavior diffusion refer to ‘peer-pressure’, i.e. normative pressure, for justifying model assumptions on behavior diffusion. However, behavioral norms are potentially of lower relevance for domestic heating behavior, because heating takes place in private, which naturally limits opportunities for normative judgment of heating practices. A promising alternative to the restriction to norms is given by Azar and Menassa (2014), capturing the role of habits.

Despite prevailing disputable theoretical foundations, the reviewed behavior diffusion studies yet partly are successfully validated. This could be explained by the existence of socio-psychological theories that justify their modeling decisions: First, Balance Theory states that people prefer consistency not just in their own atti-

study inter-linking diffusion of technology and behavior to a co-diffusion is found by this review. This suggests that an agent-based model on co-diffusion of technology and behavior concerning heating energy efficiency might not exist yet. However, modeling studies on diffusion of technology and behavior, which partly meet the stated requirements, do exist (see Table 1). Regarding the stated criteria, in both fields of application, modeling studies can be found which represent households with agents. Lifestyles and social theory are applied to less extent in the modeling studies on diffusions of behavior. Even though none of these modeling studies explicitly represents heating behavior, all refer to energy and resource conservation behavior. Thus, we see potential to apply these to heating behavior.

Table 1: Reviewed Modeling Studies Compared to Design Requirements (see text). Signs +, - and ~ Represent ‘meets requirement’, ‘does not meet requirement’ and ‘meets to requirement to some extent’, Respectively

Authors	ABM	Household	Lifestyle	Social Netw.	Theory	Validation	Co-Diff.
Zhang et al. (2011)	+	-	-	+	+/-	-	-
Azar and Menassa (2014; 2010; 2011)	+	-	-	+	+/-	-	-
Chen et al. (2012)	+	+	-	+	+/-	+	-
Anderson et al. (2014)	+	-	-	+	+/-	-	-
Schwarz et al. (2009)	+	+	+	+	+	+	-
Sopha et al. (2013)	+	+	+	+	+	+	-
Kroh et al. (2012)	+	+	+	+	+	+	-
Zhang and Nuttal (2007)	+	+	-	+	+	-	-
Jager and Janssen (2012)	+	~	~	+	+	~	~

tudes, motivations and behaviors, but also in their interpersonal relationships. This implies that consumers change their behavior if it is not ‘in balance’ with their peers. Second, Social Learning Theory states that peers are a source for constant learning, implying that consumers do reproduce their peers’ heating behavior (Jackson 2005). Both behavioral theories are supported by the findings of Peschiera et al. (2010) and, furthermore, justify the reviewed diffusion models. Most importantly, they underscore the relevance of behavior diffusion in social networks. As maybe the only major shortcoming, behavioral relapse after interventions, as observed by Peschiera et al. (2010), is not captured by the reviewed agent-based modeling studies. This could be regarded as a blind spot.

We draw three conclusions concerning applied diffusion theories. First, we regard Theory of Planned Behavior appropriate for describing technology diffusion. Second, Balance Theory and Social Learning Theory complement and support modeling assumptions on reviewed behavior diffusion models. Third, the concept of behavior relapse should be included into future models of behavior diffusion.

Conclusions on Reviewed Models.

None of the reviewed modeling studies meets all of the above stated requirements. In other words, no modeling

DISCUSSION: LESSONS LEARNED FOR BUILDING A DECISION SUPPORT MODEL

So far, we have explored simulation studies and theories that could be used to model effects of Transformational Products on a societal scale. Stressing that a combination of these models could meet all requirements, we are briefly outlining the design of such a joint simulation model. To achieve this, TAPAS validation provides an accepted methodological basis for building a new model on the basis of an existing one. This way, even the validation properties of an initial agent-based models could partially be transferred to a joint model (Frenken 2004; Windrum et al. 2007).

Such a model could contain two modules, which combined could meet all given criteria. First, technology diffusion of Transformational Products could be modeled for instance based upon the technology diffusion model by Schwarz (2007). Second, behavior diffusion could be modeled on the basis of e.g. Azar and Menassa (2014). Finally, a link between these modules would be established. This link would capture the effect of Transformational Products on heating behavior in households.

Even though linking existing agent-based models is possible, we expect it not to be trivial. A challenge will be matching them in degree of aggregation. E.g. scales

in existing models differ for aggregation of agents: in most of the reviewed technology diffusion studies these are ‘super-households’, representing numerous entities. In contrast, behavior diffusion is modeled for individual persons, i.e. small households. A matching scale will have to be found when joining the two fields.

Having outlined a possible agent-based model on the effect of Transformational Products, we are describing how such a model could assist policy makers. Co-diffusion of this technology and the behavior it supports, though being a complex process, structurally is a *diffusion process*. Thus, existing knowledge on policy options to diffusion (see Tao Zhang and Nuttall 2007) can potentially be transferred to it. However, future research has to show if this combined diffusion leads to qualitative difference or if transferability stays intact.

In combination with knowledge on sensitivity of diffusion to policy, the proposed simulation model can be put to use in several ways: first, ideal policy mixes could be identified to manage the diffusion process ‘towards’ multiple aims, e.g. energy efficiency, social justice (see Chawla and Pollitt 2012) and ease of policy implementation. Second, insight could be gained into possible futures of diffusion of Transformational Products. Third, a simulation model would enable us to suggest empirical research for filling key knowledge gaps. Fourth, the proposed agent-based model would be a theoretical contribution to the understanding of co-diffusion of technology and behavior. These applications can be particularly meaningful when tailored to a case area. Proposed policy mixes can be fitted to local actors and their competences. Potential effects of Transformational Product could be differentiated for district or street level. Finally, case studies would tie links to empirical research and support empirical studies.

In this setting, the proposed agent-based diffusion model would structurally be capable of answering, among others, the following questions: (1) Which market introduction strategy would reach which social groups, e.g. such that are prone to energy vulnerability? (2) Regarding the expected potential of Transformational Products in changing routinized behavior and triggering of behavior diffusion: what is the performance of a marketing strategy of *lending* compared to *owning*? (3) Under which conditions can Transformational Objects, addressing behavior, exceed automatizing appliances, e.g. Smart Meters, in fostering energy efficiency? Though prone to uncertainty, these questions are particularly difficult to assess without a simulation modeling.

CONCLUSIONS

Technological interventions to domestic heating behavior, e.g. so-called Transformational Products may prove valuable for increasing domestic energy efficiency. We have discussed the literature on agent-based models on diffusion of technology and behavior, respectively, that could describe the effects of Transformational Products

in society and the potentials of policy to diffusion governance. Important modeled aspects of agent-based modeling in this respect are diffusion of (1) technology and (2) behavior. Individually, these are well established, but from the various models that have been found, none can directly tackle the co-diffusion of technology and behavior. Nevertheless, they provide ingredients. We conclude that from developing a model for Transformational Products we can expect (1) results that identify a good policy mix to governance the co-diffusion of Transformational Products and sustainable heating behavior, (2) results that enable exploration of possible futures, (3) results that support empirical research in this field, and (4) theoretical contributions to the body of knowledge on co-diffusion. These results become particularly meaningful when applied to a case study, e.g. an urban area.

FURTHER RESEARCH

Future research should extend the search for exiting co-diffusion models, e.g. to other types of simulation models than agent-based models or beyond application to the energy sector. Furthermore, when developing the proposed simulation model, it should be linked to future empirical research on Transformational Products.

ACKNOWLEDGEMENTS

We thank Georg Holtz for fruitful discussions on socio-psychological theory and Paulien Herder for useful and constructive recommendations on this paper.

REFERENCES

- Anderson, K., S. Lee, and C. Menassa. 2014. “Impact of Social Network Type and Structure on Modeling Normative Energy Use Behavior Interventions.” *Journal of Computing in Civil Engineering* 28, No.1, 30–39.
- Azar, E., and C. Menassa. 2014. “Framework to Evaluate Energy-Saving Potential from Occupancy Interventions in Typical Commercial Buildings in the United States.” *Journal of Computing in Civil Engineering* 28, No.1, 63–78.
- Azar, E., and C. Menassa. 2010. “A Conceptual Framework to Energy Estimation in Buildings Using Agent Based Modeling.” In *Proceedings of the Winter Simulation Conference*, 3145–56.
- Azar, E., and C. Menassa. 2011. “Agent-Based Modeling of Occupants and Their Impact on Energy Use in Commercial Buildings.” *Journal of Computing in Civil Engineering* 26, No.4, 506–18.
- Balaras, C.A., A.G. Gaglia, E. Georgopoulou, S. Mirasgedis, Y. Sarafidis, and D.P. Lalas. 2007. “European Residential Buildings and Empirical Assessment of the Hellenic Building Stock, Energy Consumption, Emissions and Potential Energy Savings.” *Building and Environment* 42, No.3, 1298–1314.
- Chappin, E. J. L. 2011. *Simulating Energy Transitions*. Next Generation Infrastructures Foundation.
- Chawla, M., and M.G. Pollitt. 2012. “Energy-Efficiency and Environmental Policies & Income Supplements in the UK: Their Evolution and Distributional Impact in Relation to Domestic Energy Bills”. Faculty of Economics, University of Cambridge.

- Chen, J., J.E. Taylor, and H.-H. Wei. 2012. "Modeling Building Occupant Network Energy Consumption Decision-Making: The Interplay between Network Structure and Conservation." *Energy and Buildings* 47, (Apr), 515–24.
- Ernst, A. 2014. "Using Spatially Explicit Marketing Data to Build Social Simulations." In *Empirical Agent-Based Modelling - Challenges and Solutions*, edited by Alexander Smajgl and Olivier Barreteau, 85–103. Springer New York.
- Fishbein, M., and I. Ajzen. 2005. "The Influence of Attitudes on Behavior." *The Handbook of Attitudes*, 173–222.
- Frenken, K. 2004. "History, State and Prospects of Evolutionary Models of Technical Change: A Review with Special Emphasis on Complexity Theory." *Complexity focus/Ed. J. Casti*.
- Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, et al. 2006. "A Standard Protocol for Describing Individual-Based and Agent-Based Models." *Ecological Modelling* 198, No.1-2, 115–26.
- Guerra-Santin, O., and L. Itard. 2010. "Occupants' Behaviour: Determinants and Effects on Residential Heating Consumption." *Building Research & Information* 38, No.3, 318–38.
- Holzhauser, S., F. Krebs, and A. Ernst. 2013. "Considering Baseline Homophily When Generating Spatial Social Networks for Agent-Based Modelling." *Computational and Mathematical Organization Theory* 19, No.2, 128–50.
- Jackson, T. 2005. *Motivating Sustainable Consumption: A Review of Evidence on Consumer Behaviour and Behavioural Change: A Report to the Sustainable Development Research Network*. Centre for Environmental Strategy, University of Surrey.
- Jager, W. 2003. "Breaking Bad Habits: A Dynamical Perspective on Habit Formation and Change." *Human Decision-Making and Environmental Perception—Understanding and Assisting Human Decision-Making in Real Life Settings*. University of Groningen.
- Jager, W., and M. Janssen. 2012. "An Updated Conceptual Framework for Integrated Modeling of Human Decision Making: The Consumat II." In *Paper for Workshop Complexity in the Real World@ ECCS*.
- Janssen, M.A., and W. Jager. 2002. "Stimulating Diffusion of Green Products." *Journal of Evolutionary Economics* 12, No.3, 283–306.
- Laschke, M., M. Hassenzahl, and S. Diefenbach. 2011. "Things with Attitude: Transformational Products." In *Create11 Conference*, 1–2.
- Lewin, K. 1951. "Field Theory in Social Science: Selected Theoretical Papers (Edited by Dorwin Cartwright)." Oxford, England: Harpers.
- Peschiera, G., J.E. Taylor, and Jeffrey A. Siegel. 2010. "Response-relapse Patterns of Building Occupant Electricity Consumption Following Exposure to Personal, Contextualized and Occupant Peer Network Utilization Data." *Energy and Buildings* 42, No.8, 1329–36.
- Rogers, E.M. 2010. *Diffusion of Innovations*. Simon and Schuster.
- Schwarz, N., and A. Ernst. 2009. "Agent-Based Modeling of the Diffusion of Environmental Innovations - An Empirical Approach." *Technological Forecasting and Social Change* 76, No.4, 497–511.
- Schwarz, N. 2007. "Umweltinnovationen und Lebensstile: eine raumbezogene, empirisch fundierte Multi-Agenten-Simulation". Marburg: Metropolis-Verl.
- Sopha, B.M., C.A. Klöckner, and E.G. Hertwich. 2011. "Exploring Policy Options for a Transition to Sustainable Heating System Diffusion Using an Agent-Based Simulation." *Energy Policy* 39, No.5, 2722–29.
- Sopha, B.M., C.A. Klöckner, and E.G. Hertwich. 2013. "Adoption and Diffusion of Heating Systems in Norway: Coupling Agent-Based Modeling with Empirical Research." *Environmental Innovation and Societal Transitions* 8, (Sep), 42–61.
- Thaler, R.H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth and Happiness*. Penguin Books Ltd.
- Windrum, P., A. Moneta, and G. Fagiolo. 2007. "Empirical Validation of Agent-Based Models: Alternatives and Prospects". *Journal of Artificial Societies and Social Simulation* 10, No.2, 8.
- Wood, G., and M. Newborough. 2003. "Dynamic Energy-Consumption Indicators for Domestic Appliances: Environment, Behaviour and Design." *Energy and Buildings* 35, No.8, 821-841.
- Zhang, T., P.-O. Siebers, and U. Aickelin. 2011. "Modelling Electricity Consumption in Office Buildings: An Agent Based Approach." *Energy and Buildings* 43, No.10, 2882–92.
- Zhang, T., and W.J. Nuttall. 2007. "An Agent Based Simulation of Smart Metering Technology Adoption." University of Cambridge.
- Zhang, T., and W.J. Nuttall. 2011. "Evaluating Government's Policies on Promoting Smart Metering Diffusion in Retail Electricity Markets via Agent-Based Simulation." *Journal of Product Innovation Management* 28, No.2, 169–86.

AUTHOR BIOGRAPHIES



Thorben Jensen was born in Langenhagen, Germany and went to the University Osnabrück, where he studied Applied System Science and Environmental Systems and Resource Management, including stays at the Universities of Granada, Angers and Svalbard. In 2013, he obtained his Master's degree, started his research on agent-based modeling at the Wuppertal Institute for Climate, Environment and Energy and became a PhD candidate at Delft University of Technology. His homepage can be found at <http://wupperinst.org>.



Dr. ir. Émile Chappin, born in Zoetermeer, the Netherlands, is an assistant professor at TU Delft. He specializes in agent-based modeling with a focus on sustainability, carbon and renewables policies, energy markets, and adaptation to climate change. He obtained his PhD from TU Delft – "Simulating Energy Transitions". Dr. Chappin is a senior research fellow at the Wuppertal Institute for Climate, Environment and Energy. His homepage can be found at <http://chappin.com/emile>.

SCENARIO ANALYSIS AND OPTIMIZATION APPROACH IN AIR QUALITY PLANNING: A CASE STUDY IN NORTHERN ITALY

Claudio Carnevale
Giovanna Finzi
Anna Pederzoli
Enrico Turrini
Marialuisa Volta

Department of Mechanical and Industrial Engineering
University of Brescia
Via Branze 38, 25123 Brescia, Italy
E-mail: claudio.carnevale@unibs.it

KEYWORDS

Integrated assessment modeling, Multi-objective optimization, scenario analysis, Air quality.

ABSTRACT

Secondary pollution derives from complex non-linear reactions involving precursor emissions, namely VOC, NO_x, NH₃, primary PM and SO₂. Due to difficulty to cope with this complexity, Decision Support Systems (DSSs) are key tools to support Environmental Authorities in planning cost-effective air quality policies that fulfill EU Directive 2008/50 requirements.

The objective of this work is to formalize and compare the scenario analysis and the multi-objective optimization approach for air quality planning purposes. A case study of Northern Italy is presented.

INTRODUCTION

Particulate Matter (PM) usually originates, through nonlinear phenomena, from precursor emissions (primary PM₁₀, ammonia, nitrogen oxides, sulfur dioxide and organic compound). The key problem of air quality Decision Makers is to develop suitable emission control strategies, aiming to the selection of the available technologies to limit the concentration of PM₁₀ in atmosphere.

Due to non linearities bringing to formation and accumulation of PM₁₀, it is very challenging to develop sound air quality policies. This task is even more difficult when considering at the same time air quality improvement and policy implementation cost.

In literature, the following methodologies are available to evaluate alternative emission reductions: (a) scenario analysis (Thunis et al., 2007), (b) cost-benefit analysis (Reis et al., 2005) (c) cost-effectiveness analysis (Carson et al., 2004) and (d) multiobjective analysis (Carnevale et al., 2008). Scenario analysis is performed by evaluating the effect of an emission reduction scenario on air quality, using modeling simulations. Cost-benefit analysis monetizes all costs and benefits associated to an emission scenario in a target function, searching for a solution that maximizes the objective function. Due to the fact that quantifying costs and

benefits of non material issues is strongly affected by uncertainties, the cost-effective approach has been introduced. It searches the best solution considering non monetizable objectives as constraints (non internalizing them in the optimization procedure). Multi-objective analysis selects the efficient solutions, considering all the targets regarded in the problem in an objective function, and stressing possible conflicts among them.

The multi-objective analysis has rarely been faced in literature, due to the difficulties to include the non-linear dynamics involved in PM₁₀ formation in the optimization problem. The pollution-precursor relationship can be simulated by deterministic 3D modeling systems, describing chemical and physical phenomena involved in pollutant formation and accumulation. Such models, due to their complexity, require high computational time and can not be implementable in an optimization problem, which needs thousands of model runs to find solutions. The identification of surrogate models synthesizing the relationship between the precursor emissions and PM₁₀ concentrations, therefore, can be a solution. (Carnevale et al., 2008).

In this work, scenario and multi-objective approach are applied and compared for a highly polluted region of Northern Italy, where the production of secondary PM₁₀ is significant, up to 50% and beyond (Carnevale et al., 2010).

METHODOLOGY

Scenario analysis

This is the approach mainly used nowadays to design "Plans and Programmes" at regional/local scale. Emission reduction measures (Policies) are selected on the basis of expert judgment or Source Apportionment and then they are tested through simulations of an air pollution model. This approach does not guarantee that Cost Effective measures are selected, and only allows for "ex-post evaluation" of costs and other impacts. This decision pathway can be easily interpreted in the light of the classical DPSIR (Drivers-Pressures-State-Impacts-Responses) scheme, adopted by the EU (EEA, 1999) as presented in Figure 1.

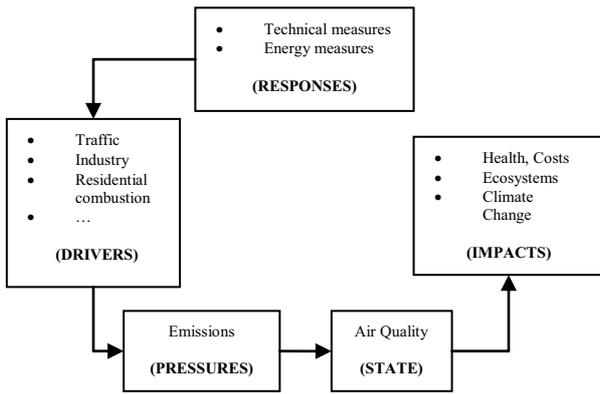


Figure 1: DPSIR Scheme for Scenario Analysis

The scenario analysis approach allows to assess the variations of the air quality indexes due to the application of a set of policies chosen a priori by the user. The problem can be formalized as follows:

$$AQI_n = f(E(\theta)) \quad \text{with } n = 1, \dots, N$$

where:

θ are the application levels of the considered technologies;

E represents the precursor emissions;

$AQI_n(\theta)$ are the Air Quality Indexes concerning different pollutants. Each Index depends on precursor emissions through emission reductions.

The decision variables θ are constrained to assume values between two extreme values:

- the CLE level, that represents the level of application for each measure as provided by european legislation for the year considered in the analysis;
- the MFR level, that is the maximum technically feasible reduction of one measure, for the year considered in the analysis.

In this approach impacts of the can be evaluated by someone that, based on its experience, acts on decision variables in order to create a more efficient scenario that can be tested again through scenario analysis.

Optimization approach

This approach, according to the DPSIR scheme, can be presented as shown in Figure 2. It faces the AQ problem defining a decision problem solved by means of optimization algorithms.

In this case the feedback from impacts is evaluated by an optimizer and, though thousands of iterations, the optimal solution is found.

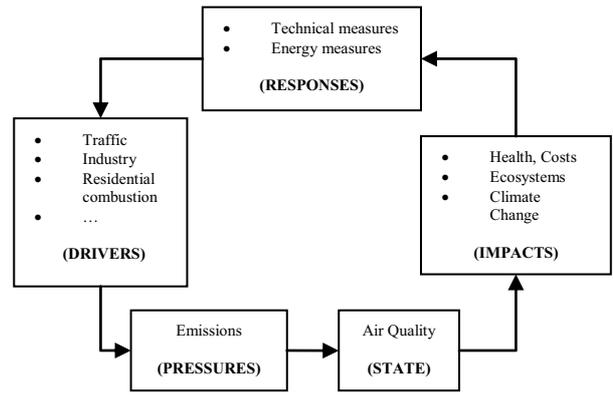


Figure 2: DPSIR Scheme for Optimization Approach

A Multi Objective problem consists of a number of objectives to be simultaneously optimized while applying a set of constraints. The problem can be formalized as follows:

$$\min J(\theta),$$

$$J(\theta): \mathbb{R}^T \rightarrow \mathbb{R}^{O_{obj}}$$

subject to:

$$\theta \in \Theta$$

where $J(\theta)$ is the objective function

T is the number of considered technologies,

O_{obj} is the number of the objectives,

θ are the decision variables constrained to assume values in the feasible decision variable set Θ .

The target of the proposed problem is to control secondary pollution at ground level. The solutions of the Multi Objective problem are the efficient emission control policies in terms of air quality and emission reduction costs. The problem can be formalized as follows:

$$\min_x J(\theta) = \min_x [AQI_n(\theta) \ C(\theta)],$$

$$\text{with } n = 1, \dots, N$$

where

- E represents the precursor emissions;
- $AQI_n(\theta)$ are (maximum N) Air Quality Indexes concerning different pollutants;
- $C(\theta)$ represents the emission reduction costs;
- θ is a vector containing the application rates of the reduction measures, constrained to be included in the feasible set Θ .

The decision problem complexity can then be reduced to a two objectives, considering a single Air Quality Index (AQI) obtained as a linear combination of the various Air Quality Indexes AQIn (plus the Cost index).

These various AQIs can be aggregated through linear combination of normalized AQIs. Finally, the previous equation can be re-written as:

$$\min_{\theta} J(\theta) = \min_{\theta} [\text{AQI}(\theta) - C(\theta)],$$

The Multi Objective optimization problem is solved following the ϵ -Constraint Method: the Air Quality objective is minimized, while the emission reduction cost objective is included in the set of constraints. In this configuration, the Multi Objective approach has the same features of the Cost Effectiveness analysis, where the Figure of Merit is

$$\min_{\theta} J(\theta) = \min_{\theta} \text{AQI}(\theta)$$

and the second objective is included in the constraints:

$$C(\theta) \leq L \quad 0 \leq L \leq \bar{L}$$

where L can assume different values in the defined range. In this way a set of effective solutions is computed and a Pareto curve can be drawn.

Air Quality objective

The Air Quality objective may consider a number of indexes related to PM10, PM2.5, ozone (eg. SOMO35, AOT40) and NOx. The case study presented in this work is focused on PM10.

All the indexes can be computed over different domains, and can be related to i.e. yearly, winter or summer periods. Starting from the local value, computed cell by cell, an aggregation function is applied, to get the scalar variable (AQI) that has to be optimized. The aggregation function can be:

- spatial Average;
- population weighted average;
- number of cells over threshold

Decision variables

The decision variables are the application rates of the emission reduction measures. In particular, two classes are considered: the end-of-pipe technologies (or technical measures) and the efficiency (or non-technical measures). Such latter measures reduce the energy consumption and as consequence the emissions. Examples of this class of measures are the behavioural changes (like the use of bicycle instead of cars for personal mobility or the reduction of temperature in buildings) or the energy saving technologies.

Applying the measures, the reduced emissions of pollutant p, due to the application of measures in sector k and activity f, are computed as follows:

$$E_{k,f,p} = \sum_{t \in T_{k,f}} (A_{k,f} \cdot \text{eff}_{k,f,t}^p) X_{k,f,t} \cdot \text{eff}_{k,f,t}^p + \sum_{t \in Z_{k,f}} (A_{k,f} \cdot \text{eff}_{k,f,t}^p) Z_{k,f,t} \cdot \text{eff}_{k,f,t}^p$$

where:

$X_{k,f,t}$: is the application rate (bounded in $[\bar{X}_{k,f,t}; \underline{X}_{k,f,t}]$) of technical measure t to sector k and activity f;

$Z_{k,f,t}$: is the application rate (bounded in $[\bar{Z}_{k,f,t}; \underline{Z}_{k,f,t}]$) of efficiency measure t to sector k and activity f;

$A_{k,f} \cdot \text{eff}_{k,f,t}^p$: is the pollutant p emission due to sector k and activity f;

$X_{k,f,t} \cdot \text{eff}_{k,f,t}^p$: is the overall technical measure t removal factor with respect to sector k, activity f and pollutant p;

$Z_{k,f,t} \cdot \text{eff}_{k,f,t}^p$: is the overall efficiency measure t removal factor with respect to sector k, activity f and pollutant p.

The total emission reduction beyond CLE scenario for a pollutant p, due to the application of a set of measures, can be calculated as the sum of the emission reductions over all the <sector-activity> pairs:

$$E_p = \sum_{k,f} E_{k,f,p}$$

Emission reduction costs

The emission reduction costs are calculated first for each sector-activity:

$$C_{k,f} = \sum_{t \in T_{k,f}} c_{k,f,t} \cdot A_{k,f} \cdot X_{k,f,t} + \sum_{t \in Z_{k,f}} c_{k,f,t} \cdot A_{k,f} \cdot Z_{k,f,t}$$

where:

$c_{k,f,t}$ is the unit cost [M€/year] for sector, activity, technology k,f,t;

$C_{k,f}$ is the total cost [M€/year] for sector, activity k,f;

$T_{k,f}$ are the technologies that can be applied in a defined sector activity.

Then, the total emission reduction cost [M€/year] is computed as:

$$C = \sum_{k,f} C_{k,f}$$

Constraints

The first constraint concerns the emission reduction cost, which cannot be greater than the available budget L.

The following constraints hold for *technical measures*.

When the substitution of old technologies is admitted, the following constraints are applied:

- to ensure the application feasibility:

$$0 \leq X_{k,f,t} \leq \bar{X}_{k,f,t} \quad \forall k \in K, f \in F_k, t \in T_{k,f};$$

- to ensure the mutual exclusion of technical measures application (for each activity and each primary pollutant, i.e. for each activity and each precursor):

$$\sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t} \leq 1 \quad \forall k \in K, f \in F_k, p \in P;$$

- to ensure that the emission reduction achieved according to the optimal solution are at least those guaranteed by the application of the technologies imposed by the Current LEgislation (for each activity and each primary pollutant):

$$\sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t} \cdot \text{eff}_{k,f,t}^p \geq \sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t}^{\text{CLE}} \cdot \text{eff}_{k,f,t}^p$$

$$\forall k \in K, f \in F_k, p \in P;$$

- to ensure that the emissions controlled according to the optimal solution are at least those controlled applying the technologies at the lower bounds imposed by the Current LEgislation:

$$\sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t} \geq \sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t}^{\text{CLE}} \quad \forall k \in K, f \in F_k, p \in P;$$

Concerning *efficiency measures*:

- to ensure the application feasibility:

$$Z_{k,f,t}^{\text{CLE}} \leq Z_{k,f,t} \leq \bar{Z}_{k,f,t} \quad \forall k \in K, f \in F_k, t \in NT_{k,f};$$

Moreover, when both technical and efficiency measures are applied, the global conservation of mass constraints have to be stated explicitly (for each activity and each primary pollutant):

$$\sum_{t \in T_{k,f}: \text{eff}_{kft}^p > 0} X_{k,f,t} \text{eff}_{kft}^p + \sum_{t \in NT_{k,f}: \text{eff}_{kft}^p > 0} Z_{k,f,t} \text{eff}_{kft}^p \leq 1$$

$$\forall k \in K, f \in F_k, p \in P$$

TEST APPLICATION RESULTS

Case study

In these section, the proposed approaches are applied and compared to the test case of Lombardia region in Northern Italy. This is one of the most polluted regions in Europe due to three main factors: high level of emissions, stagnant meteorological conditions (low wind speed and temperature inversions) and a complex topography that prevents access to strong winds. For these reasons, unless the European legislation is applied, high levels of particulate matter are still a major concern in the region. The geographical domain

was discretized with a 6 x 6 km² grid and comprises roughly 6000 cells (see Figure 3).

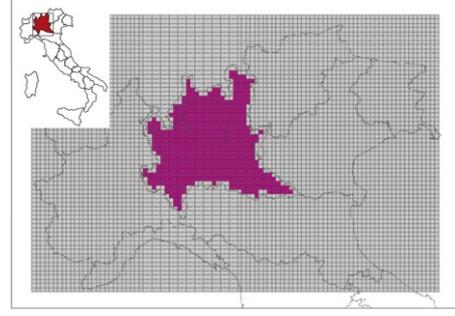


Figure 3: Lombardia region Domain.

The air quality index (AQI) is the yearly average of PM10. The relationship between such index and the decion variables, namely the annual emissions of the precursors (NOx, VOC, NH3, PM10, PM2.5, SO2) for each domain cell, is modelled by Artificial Neural Networks (ANNs). The ANNs are identified processing long-term simulations of TCAM model. Such simulations are selected assessing of nonlinear relationship between the precursor emissions and PM concentrations. Such analysis has been performed implementing the Factor Separation Analysis (Canevale et al., 2010) and has produced 20 scenarios varing emissions between CLE2010 and MFR2020.

A quadrant shape input configuration has been used, as shown in the Figure 3:

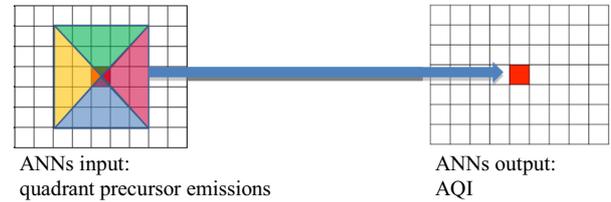


Figure 3: Quadrant Shape input Configuration.

This shape of input allows considering the prevalent wind directions over the domain: the North-South direction follows the Po Valley axes and the East-West direction is the breezes axes. So to simulate the Air Quality Index on a particular cell, 48 input data are considered:

- the 6 emissions precursor under study (NOx, VOC, NH3, PM10, PM25 and SO2);
- the 4 quadrants;
- the 2 emission levels: *low* for areal emissions and *high* for point sources.

The dimension of the input quadrants is 24km.

The source-receptor models are Feed Forward Artificial Neural Networks, with one hidden layer. To select, the best ANN structure, the following tests have been performed:

- number of neurons of the hidden layer: 10 or

20;

- transfer functions of the first and hidden layer: linear, tangent-sigmoid, logarithmic sigmoid;
- number of epochs (100, 200, 300, 400);

The identification dataset contains the 80% of the TCAM simulation cells, while the 20% of the cells (spatially uniformly distributed) is kept for validation. The ANN structure with lower Mean Squared Error is selected and used in the next phase of the work.

Traffic Scenario (TS) analysis

An emission reduction scenario has been performed considering the application of the new EURO standard to the all vehicles (EURO V and VI) substituting the older standars. And, in addition to this, the application at the maximum possible level, of three efficiency measures (efficiency measures):

- bus investment;
- construction of new bicycle paths;
- lowered speed on highways.

The simulation of this scenario has been performed using the RIAT+ tool (Carnevale et al. in press) and shows that, starting from a CLE 2010 scenario with a PM10 yearly average of $27.3 \mu\text{g}/\text{m}^3$, the application of these technologies would cost 170M€, allowing a mean reduction of 6% in the PM10 average concentrations over the domain. A reduction of 6% in health costs, due to the months of life lost, has been estimated using ExternE approach (Bickel et al. 2005).

Table 1: Traffic Scenario (TS) features.

Impacts	CLE	TS
Emission reduction costs [M€/year]	0	170
PM10 [$\mu\text{g}/\text{m}^3$]	27.3	- 6%
Health costs (due to months of life lost) [€]		- 6%

In Figures 4 and 5, the spatial distribution of the yearly PM10 concentrations is shown for CLE2010 and TS. It is clear that the latter scenario is reducing PM10 particularly in the central most populated and industrialized cells between the cities of Milano, Bergamo and Brescia.

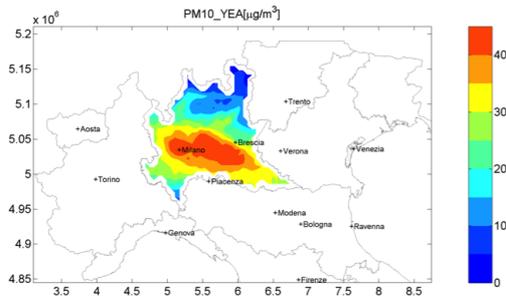


Figure 4: PM10 yearly average [$\mu\text{g}/\text{m}^3$] map for CLE2010.

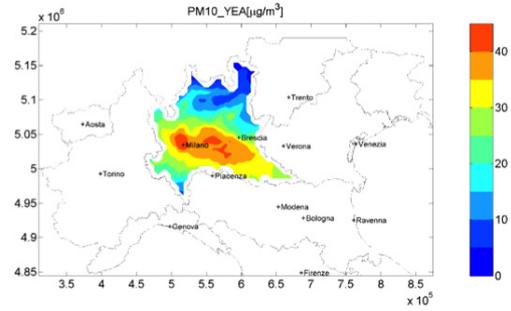


Figure 5: PM10 yearly average [$\mu\text{g}/\text{m}^3$] map for TS.

Figure 6 shows the emission reductions in each macrosector for TS. The selected measures are reducing more than 35000 Kton/year of NOx emissions and around 10000 Kton/year of VOC emissions.

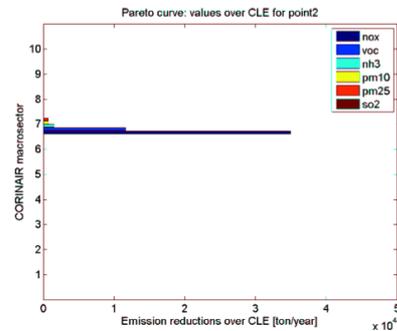


Figure 6: Emission Reductions [ton/year] in each Macrosector for TS.

Optimization approach

Applying a Cost-Effectiveness analysis at the same cost of the Traffic Scenario (170M€), an Optimized Scenario (OS) has been computed. It represents the most performing scenario applying the most effective measures. Both the PM10 mean concentrations and the health costs have a significant reduction, going respectively from 6% to 21% and from 6% to 19%, as shown in summary Table 2.

Table 2: Traffic Scenario (TS) and Optimized Scenario (OS).

Impacts	CLE	TS	OS
Emission reduction costs [M€/year]	0	170	170
PM10 [mg/m3]	27.3	- 6%	- 21%
Health costs (due to months of life lost) [€]		- 6%	- 19%

Figure 7 depicts the Pareto curve (emission reduction cost objective vs. mean PM10 concentrations) that results from the Multi Objective optimization. Starting from CLE scenario, the curve shows the optimal

solutions at different costs. In particular two points are highlighted: Traffic Scenario (green triangle) and the Optimized Scenario (red square).

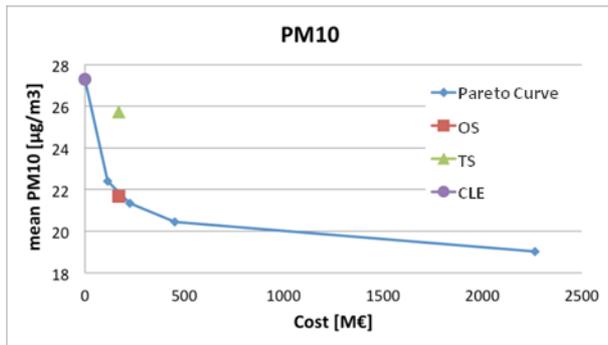


Figure 7: Pareto curve, TS (green triangle) and OS (red square).

Figure 8 shows the map of the yearly PM10 concentrations for OS. The highest concentrations are essentially disappeared over the industrialized area between Milano and Brescia.

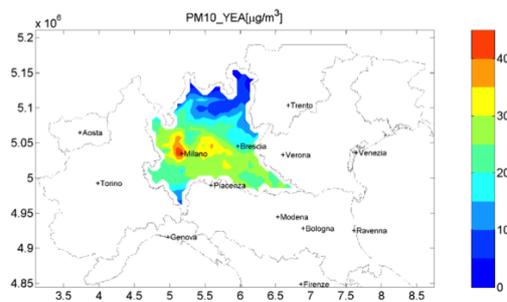


Figure 8: PM10 yearly average $[\mu\text{g}/\text{m}^3]$ map for the optimized scenario.

Figure 9 shows the costs in each macrosector for OS. More than 140M€ over the total (170M€) are allocated for macrosector 10 (Agriculture). Macrosector 7 (Transports) and 2 (Non industrial combustion) are relevant. Figure 10 shows the emission reductions in each macrosector for the same scenario. The measures for Agriculture allow to reduce a great amount of NH3 emissions, but also the limited budget invested in macrosectors 7 and 8 (Transports and Other Mobile Sources) allows to reduce great amounts of emissions, in particular NOX emissions.

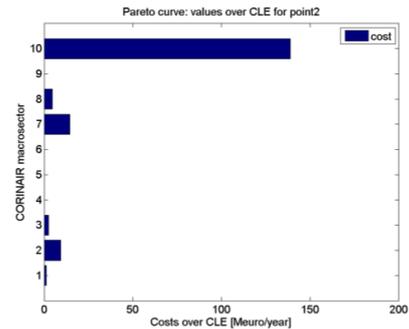


Figure 9: Costs [M€/year] in each Macrosector for OS.

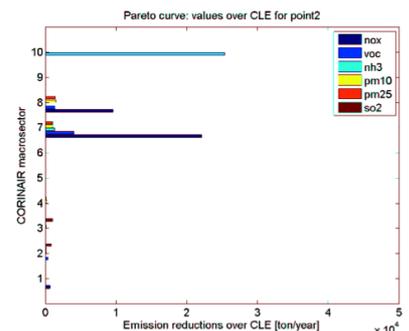


Figure 10: Emission Reductions [ton/year] in each Macrosector for OS.

CONCLUSIONS

In this paper the comparison between two approaches for air quality planning is presented. The first one is the scenario approach; it allows to assess the variations of the air quality indexes due to the application of a set of policies chosen by the user. Since the possible technological or non technological measures that can be implemented to reduce air pollution are hundreds, this approach does not guarantee that the most efficient combination of measures is identified, even though a large number of scenarios are assessed.

The Multi Objective approach optimizes a number of objectives simultaneously while applying a set of constraints. It allows to find the most efficient set of measures that guarantees to achieve the higher reduction of secondary pollution over the domain, at minimum costs.

The case study presented shows that the scenario analysis focused the traffic emission macrosector is not efficient. The optimization approach, taking into account hundreds of different measures, is able to find air quality policies that are more effective on air quality, and consequently on health effects, and emission reduction costs.

REFERENCES

Bickel, P.; Friedrich, R.; 2005. "ExternE: externalities of energy, methodology 2005 Update." *Tech. rep. IER*, University of Stuttgart.

- Carnevale, C.; Finzi, G.; Pisoni, E. and Volta, M.; 2008. "Modelling assessment of PM10 exposure control policies in Northern Italy." *Ecological Modelling* 217, 219-229.
- Carnevale C.; Finzi G.; Pisoni E. and Volta M.; 2009. "Neuro-fuzzy and neural network systems for air quality control." *Atmospheric Environment*, Volume 43, 4811-4821.
- Carnevale, C.; Pisoni, E.; and Volta, M.; 2010. "A non-linear analysis to detect the origin of PM10 concentrations in Northern Italy." *Science of the Total Environment* 409, 182-191.
- Carnevale, C.; Finzi, G.; Pederzoli, A.; Turrini, E.; Volta, M.; Guariso, G.; Gianfreda, R.; Maffei, G.; Pisoni, E.; Thunis, P.; Markl-Hummel, L.; Perron, G.; Blond, N.; Weber, C.; Clappier, A.; Dunardin, V.; in press. "Exploring trade-offs between air pollutants through an Integrated Assessment Model".
- Carlson, D.; Haurie, A.; Vial, J.P.; and Zachary, D.; 2004. "Large-scale convex optimization methods for air quality policy assessment." *Automatica*, 40, 385-395
- European Environment Agency, 1999. *Environmental indicators: Typology and overview*. Technical report No 25/1999, EEA Copenhagen, (Sep)
- Janssen, S.; Ewert, F.; Li, Hongtao; Athanasiadis, I.N.; Wien, J.J.F.; Théron, O.; Knapen, M.J.R.; Bezlepkina, I.; Alkan Olsson, J.; Rizzoli, A.E.; Belhouchette, H.; Svensson, M. and Van Ittersum, M.K.; 2009. "Defining assessment projects and scenarios for policy support: use of ontology in integrated assessment and modelling." *Environmental Modelling & Software* 24, 1491-1500.
- Reis, S.; Nitter, S.; Friedrich, R.; 2005. "Innovative approaches in integrated assessment modelling of European air pollution control strategies – Implications of dealing with multi-pollutant multi-effect problems." *Environmental Modelling and Software*, 20, 1524-1531.
- Thunis, P.; Rouil, L.; Cuvelier, C.; Stern, R.; Kerschbaumer, A.; Bessagnet, B.; Schaap, M.; Builtjes, P.; Tarrason, L.; Douros, J.; Moussiopoulos, N.; Pirovano, G.; Bedogni, M.; 2007. "Analysis of model responses to emission-reduction scenarios within the CityDelta project." *Atmospheric Environment*, 41(1), 208-220.

AUTHOR BIOGRAPHIES

CLAUDIO CARNEVALE is Assistant Professor at University of Brescia since 2006. He holds a Ph.D. in Information Engineering (University of Brescia) and he has been involved in a number of National and International projects (QUITSAT, RIAT, OPERA,

APPRAISAL). He is author or co-author of around 80 scientific papers.

E-mail: claudio.carnevale@unibs.it

GIOVANNA FINZI is Full professor in Environmental Modelling and coordinates the ESMA research team, working on environmental decision support systems. She has been national delegate in the MC of several COST Actions, in EMEP TFMM and in TF-HTAP (UNECE LRTAP Convention) and member of IFAC, AGU, IEEE. She is in the Steering Group of the APPRAISAL project. She co-authored several peer-reviewed scientific papers.

E-mail: giovanna.finzi@unibs.it

ANNA PEDERZOLI is currently working at DIMI, University of Brescia. Her main research interests concern atmospheric physics and chemistry and the evaluation of the impact of air quality control policies. She is involved in a number of national and international projects.

E-mail: anna.pederzoli@unibs.it

ENRICO TURRINI is Research Fellow and Ph.D. student in Technology for Health at University of Brescia since 2013. He has been involved in a number of national and international projects concerning optimal policy selection.

E-mail: enrico.turrini@unibs.it

MARIALUISA VOLTA is Associate Professor in Environmental Modelling, with a Ph.D. in Information Engineering. She has been involved in several national and European Projects, in particular as Project Leader of RIAT JRC project, chair of the OPERA LIFE+ project Steering Group, and coordinator of the APPRAISAL FP7 project. She is a national delegate at UNECE-TFIAM, a member of NIAM and of IFAC Technical Committee on Modelling and Control of Environmental Systems. She is author or co-author of around 120 scientific papers (65 peer-reviewed papers).

E-mail: marialuisa.volta@unibs.it

TACTICAL VERSUS OPERATIONAL DISCRETE EVENT SIMULATION: A BREAST SCREENING CASE STUDY

Andrea Lodi and Paolo Tubertini
DEI - Università di Bologna

Viale Risorgimento 2 - 40136
Bologna, Italy
E-mail: {andrea.lodi,
paolo.tubertini}@unibo.it

Roberto Grilli and Francesca Senese
Agenzia Sanitaria e Sociale
dell'Emilia-Romagna
Viale A. Moro, 21 - 40127
Bologna, Italy
E-mail: {RGrilli, FSenese}
@regione.emilia-romagna.it

KEYWORDS

Discrete Event Simulation, Capacity Planning, Policy Modelling.

ABSTRACT

The Regional Agency for Health and Social Care, (ASSR) of Emilia Romagna, a regional research center for innovation and improvement, is trying to understand how Discrete Event Simulation (DES) technique can be incorporated to support decision making in the field of health services. Most of decision support systems used by regional and local health planners face problems from the epidemiological point of view whereas few studies are focused on dynamic capacity planning.

The aim of this work is to study how two different DES software packages can be effectively applied to support tactical and operational decision-making processes. In order to validate our DES model we built a breast-screening pathway and we showed how this model could predict lead time performances with different capacity settings. We focused our work on the inclusion in 2010 breast screening program of 45-49 and 70-74 women age bands, by showing how simulation can help to understand resource re-sizing needs as a consequence of an increase in the demand of services. We analyzed the breast program with two DES software packages in order to show how different tools can help stakeholders that operate at different decision levels (local vs regional). Both simulation models were tested over a one year time horizon, we distinguish the operational from the tactical level as a consequence of the different potential degree of detail that we can reach in the care pathway modeling.

INTRODUCTION

Since 1996, the Department of Public Health of the Emilia-Romagna region, based on national and international scientific community recommendations, provides a free screening program for early detection of breast cancer. The screening program offers scheduled checks to women, residents or domiciled in the region, falling in those age groups in which the risk of cancer increases the effectiveness of early diagnosis and appropriate treatment, reducing the risk of death. The monitoring is done through periodic checkups, namely mammography tests performed every two years.

Local Health Units, which are coordinated and supervised by the Emilia-Romagna Regional Health Authority, are in charge of managing screening programs. The program is characterized by an integrated diagnostic-therapeutic pathway that follows the patient from the screening test up to surgery treatment and follow-up treatments.

Until 2010 women in 50-69 age bands were the target population, in 2010 the Regional health Authority extended the breast screening program to the 45-49 and 70-74 age bands.

LITERATURE REVIEW

In the literature, breast-screening programs have been studied with simulation techniques considering several aspects. In (Michaelson et al. 1999) the authors use biologically-based data from the literature on the rates of tumor growth and spread, to calculate the course of breast cancer growth and metastasis aiming at defining the optimal screening interval for detection of breast cancer prior to distant metastatic spread. In (Fryback et al. 2006) the authors focus their attention on epidemiological aspects by simulating 25 years of U.S. women population evolution and addressing what-if questions about effectiveness of screening and treatment protocols, as well as estimating benefits to women of specific ages and screening histories. Improving health outcomes through effective diagnostic and treatment is certainly the overriding objective of screening programs, nevertheless an in-depth evaluation of resource consumption and financial sustainability is very important guarantee the success of the program. In (Brown and Fintor 1993) and (Hunter et al. 2004) simulation studies evaluate the cost-effectiveness of breast screening policies testing different scenarios that concern epidemiological trends, population age bands inclusion and possible outcomes and outputs both in terms of quality of care and of financial impact. In this context the MISCAN (MISrosimulation SCreening ANalysis) model, which uses Monte Carlo micro-simulation of a large number of life histories according to the epidemiology of the disease in question, has been used to model and test various breast screening issues in Italy (Paci et al. 1995), Germany (Beemsterboer et al. 1994) and Australia (Carter et al. 1993). Surveys on cost effectiveness models regarding breast screening programs can be found in

(Brown and Finton 1993) and in (Fone et al. 2003). In addition to epidemiological and cost-effectiveness analysis some side aspects regarding screening programs have been investigated in (Brailsford and Schmidt 2003) such as the impact of patients behavior on attendance rates. In conclusion, screening programs are characterized by a set of guidelines and benchmarks that are used for performance monitoring. The quality of service can be evaluated according to health goals and organizational objectives. Indicators such as stage at diagnosis (defined as the ability to anticipate the detection of cancer), quality of care (defined as the reduction of diagnostic errors) and 5-year survival rate after surgery treatment are primary objectives for health managers. Nevertheless, screening programs have to be evaluated with an organizational and managerial perspective since, in order to provide services, a set of facilities and associated resources have to be identified. Future volume of activities and their financial sustainability as well as resources availability and waiting times can affect treatment effectiveness and are usually monitored and taken into consideration during planning activities.

BREAST SCREENING: PLANNING PROBLEM

As previously stated, in 2010 Emilia-Romagna Regional Health system decided to extend the breast screening program to 45-49 and 70-74 age bands. Each Local Health Authority, supported by the regional one, had to decide what resources should be resized even if at the time of the planning process it was not clear the impact of the extension of the screening coverage in terms of waiting time and lead time performances. The aim of this work is to study how two different DES software packages could have been effectively applied to support the process from a tactical and operational point of view. The case study is only focused on capacity planning since cost effectiveness analysis as well as screening frequency policies were already defined by strategic planners.

Breast Screening Pathway

Screening program can be generically described as a care pathway. In (Naldoni et al. 2012) Emilia-Romagna regional guidelines and benchmarks are reported. It is possible to split the breast screening pathway in three main components: the first contact and appointment management, the first level examination and the second level examinations.

Invitation, appointment reschedule and reminder.

Each woman falling in the target age band is invited every two years to undergo a screening test. The woman is invited by the Local Screening Center that communicates day and time of the appointment. If the candidate is unable to attend she can reschedule the appointment, otherwise she goes directly to the diagnostic ambulatory. In case of no shows, the candidate is contacted and a new appointment is planned.

Contact activities are managed by the Screening Center, an organizational unit that works as an interface between screening candidates and program activities. This structure is responsible of breast, uterine and colon-rectum screening programs, therefore its operators are shared resources and segment their weekly activity in dedicated time slots for each program.

The Screening Center is responsible of monthly supervision of invitations and no shows, appointment reschedule and monthly reminder letter management (see Figure 1 for pathway representation).

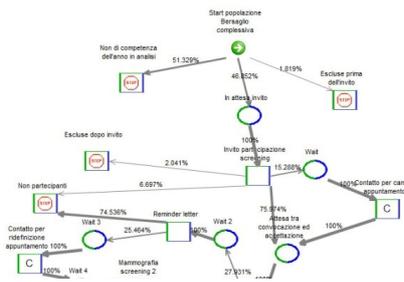


Figure 1: Screening Pathway: Invitation Activities

First level examination

The day of the appointment the candidate undergoes a mammography test that is performed by a radiologist technician by using a Breast Computed Radiography Scanner. After the screening test is performed, the recorded images are sent in a digital way to Radiology senology specialists. When available a specialist analyzes the test and gives his/her diagnosis that can be negative, positive or uncertain. A test has to be analyzed at least by two different physicians and if both of them consider it negative, the patient will receive a letter confirming that no evidence of potential cancer was found. If at least one of the diagnosis is uncertain, a third physician analyzes the test and decides if the patient must undergo in-depth examinations.

Even though the examinations are carried out at a local level, the analysis can be done in real time anywhere, since the images are remotely available in a digital format. In the ideal case it is possible that on the same day of the examination all diagnostic analysis are performed (see Figure 2 for pathway representation).

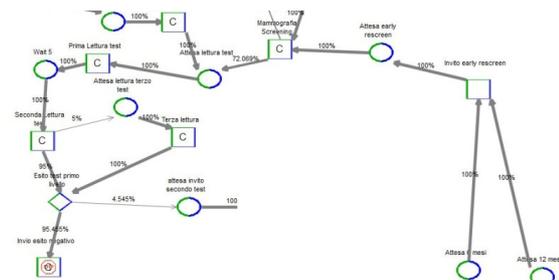


Figure 2: Screening Pathway: First Level Activities

Second level examination

In depth examinations are performed for positive or uncertain patients. After a clinical examination the radiologist decides, depending on first level test results, if the patient should undergo detailed mammography, ultrasound or magnetic resonance imaging (MRI) examinations. If non-invasive examinations show the potential presence of a cancer, before proceeding with surgical activities, an invasive test such as cytology or micro-biopsy, is done in order to ascertain the presence of a tumor (see Figure 3 for pathway representation).

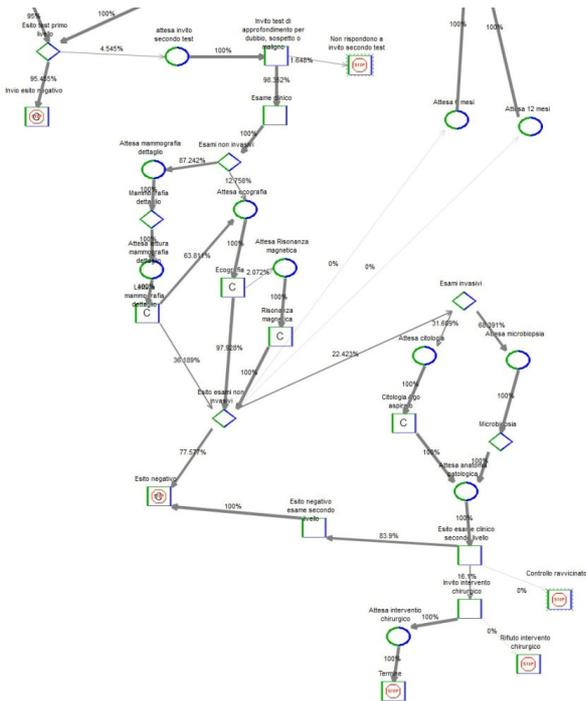


Figure 3: Screening Pathway: Second Level Activities

Key Performance Indexes

In order to monitor the organizational performances of the screening program two lead times are monitored, the time elapsed from mammography test to the dispatch of the letter with the first level negative result and the time elapsed from the mammography test to the first in-depth appointment. For each KPI two thresholds are monitored, first-level letter dispatch within 15 and 21 days and the first in-depth appointment within 21 and 28 days. Both lead time indicators measure Radiology senology specialists and Screening Center performances.

Local Health Authority data

Tactical versus operational planning has to take into account different levels of detail during the data collection process. Since screening programs are organized at a local level, we focused our case study on a regional local health authority. We collected data regarding volume of activities and resources involved in 2009 in order to build and validate a DES model. In 2009 the target population (women in the 50-69 age band) was equal to 51,462 residents and 24,111 of them

were invited to attend the screening test. The 65.77% of invited women attended the test in the planned day, 15.29% called for a reschedule, 27.93% did not attend to the first appointment and 74.54% of them did not answer to the reminder letter. Two screening center operators managed appointment agenda and need three minutes for each call.

Screening tests were organized in eight public clinics (LOC) spread over the local health authority territory.

Table 1: First Level Examination Resources Available in 2009

Resource ID	Number of working days per week	Number of mammograms per day	Working months per year
LOC 1	2	24	10
LOC 2	2	24	10
LOC 3	4	29	10
LOC 4	2	24	10
LOC 5	4	24	10
LOC 6	3	24	10
LOC 7	2	24	10
LOC 8	4	54	11

Table 1 shows the yearly schedule of each clinic. First, analysis were partially outsourced, in 2009 30% of them were managed by an Hospital trust that guaranteed diagnostic results within 5 working days.

Seven Radiology senology specialists (RSS) working for the Local Health Authority managed the remaining 70% first analysis as well as all second and third ones (see Table 2). Monday, Tuesday and Thursday were dedicated to first level analysis and each mammogram test analysis took on average three minutes.

Screening Center manager dedicated on each day one hour and a half to send 140 negative result letters. In-depth examination were managed by Local Health Authority Radiology senology specialists by following a fixed schedule, on Wednesday non-invasive examinations were performed invasive ones were planned on Friday.

Table 2: Radiology Senology Specialists Available in 2009

Specialist ID	Months of activity (included holidays)	Number holidays during activity months (days)
RSS 1	12	20
RSS 2	10	40
RSS 3	3	10
RSS 4	12	52
RSS 5	3	17
RSS 6	4	23
RSS 7	0	365

Table 3 shows the workload of diagnostic sonography tests, magnetic resonance imaging tests, detailed mammogram procedures, micro biopsies and cytology test reading in terms of volume of activities and time required by each activity.

Table 3: Second Level Examination 2009 Workload

Activity type	Number	Mean time (minutes)
Detailed mammogram procedure	677	15
Magnetic Resonance Imaging	373	20
Diagnostic sonography	373	15
Micro biopsies	119	30
Cytology test reading	55	30

In addition to screening-driven activities, radiologists had to deliver a set of services associated with regular outpatient activities within the public sector (see Table 4).

Table 4: Non-Screening Activities in 2009

Activity type	Number	Mean time (minutes)
Computed Radiography	12,379	10
Magnetic Resonance Imaging	762	30
Diagnostic sonography	2,276	20
X-ray computed tomography	863	20

As a result, in 2009 81.40% of first level negative results were sent within 15 days and 89.20% within 21 days, whereas 78.14% of first in-depth examinations were performed within 21 days and 84.88% within 28 days.

SIMULATION RESULTS

We implemented an operational model and a tactical one and we validated them on 2009 data. We recall that both simulation models have been tested over a one year time horizon and we distinguish the operational from the tactical level as a consequence of the different potential degree of detail that we can reach in the care pathway modeling. Namely, the operational model, implemented with Simul8, describes the daily organization of the resources and entails the possibility of analyzing the performances of the care pathway on a weekly or monthly basis. Conversely, the model implemented with Scenario Generator can only be used to support tactical decision making processes because it only allows for a weekly description of the system resources.

Below we present the results for both models.

Simul8 (Operational level) model

We developed an operational model using Simul8 (Concannon et al. 2007), a general purpose DES software, defining Radiology senology specialists, public clinics and Screening Center detailed schedules. Then, we tested the program extension impact under different system configurations. The target population in 2010 increased up to 80,289 women in 49-70 age band where 48,165 had to be invited.

As a first planning hypothesis we fixed the rates of (i) invited women that attend the test in the planned day, (ii) no-shows after first invitation, (iii) appointment reschedules and (iv) no-shows after reminder letter (see Table 5 for detailed forecasted activity volumes). Then, we tested the performance worsening in case of no Radiology senology specialists resizing and by considering that in 2010 the outsourcing contract was expired. As it is clear in Table 2, the real number of active Radiology senology specialists in 2009 was less than the theoretical one. We then tested the system behavior under the hypothesis of five Radiology senology specialists working full time for the Local Health Authority. That hypothesis holds since one of the seven physicians left in the first days of 2009 and the second one would have been pregnant during 2010. Observing past annual volume of holidays, we identified an average of 49 days off per year per physician. The proposed setting would have led, for 2010, to a 67.01% of first level negative results sent within 15 days and 73.25% within 21 days, while 63.04% of first in depth examination would have been performed within 21 days and 69.43% within 28 days.

As a second planning hypothesis, we considered the impact of increasing the number of Radiology senology specialists available until near optimal performances were reached for both KPIs. It is important to say that the result is strongly influenced by the policy implemented for holidays. It is possible to reach a 98.76% of first level negative results sent within 15 days and 99.12% within 21 days with just one additional resource if physicians holidays never overlap. This would not have been the case even with two additional resources if holidays overlap. Two additional resources would have led just to a 76.94 % of first level negative results sent within 15 days and 76.94 % within 21 days.

Since 2010 data about real performances are available, we decided to test how the resizing proposed by the model considering 2009 population behavior would have been able to cope with real 2010 population behavior (see Table 5).

Table 5: 2010 Forecasted and Real Volume of Activities

Activity type	2010 Forecast	2010 Real activities
Accepted after first invitation	31,679	28,909
No shows	13,453	13,087
Examination after reminder letter	3,426	2,005
First level examination	35,104	30,914
Negative examination	33,509	29,473
Positive examination	1,595	1,430
Not responding to in depth examination	24	11
Responding to in depth examination	1,572	1,419
Detailed mammographies (DM)	1,371	1,238
Diagnostic sonographies	201	181
Diagnostic sonographies after DM	875	1,267
MRI tests	22	10
Invasive examinations	319	288

The proposed setting would have led to a 97.93% of first level negative results sent within 15 days and 99.19% within 21 days, while 98.77% of first in-depth examinations would have been performed within 21 days and 99.13% within 28 days. These result show the proposed resizing would have been effective on the population behavior for 2010.

In addition to screening activities Radiology senology specialists are also involved in general outpatients activities concerning detailed mammography, ultrasound or MRI examinations. In 2010 an increase in the demand of non-screening services was recorded (see Table 6) and we tested how that could have impacted on our proposed resource resizing.

Table 6: Recorded Non-Screening Activities in 2010

Activity type	Number	Mean time (minutes)
Computed Radiography	19,680	10
Magnetic Resonance Imaging	413	30
Diagnostic sonography	8,003	20
X-ray comp. tomography	1,536	20

The proposed setting would have led to a 13.75% of first level negative results sent within 15 days and 16.76% within 21 days while 14.48% of first in depth examination were performed within 21 days and 16.94% within 28 days, i.e., a dramatic worsening in performance. To face the increased volume of non-screening activities two additional Radiology senology

specialists should have been included by the Local Health Authority.

Scenario Generator (Tactical level) model

In the previous section we analyzed how an operational model could have supported Local Health Authority planning. At a regional level one could be tempted use less detailed information of the system at hands to do a more tactical planning, for example without resource daily schedules. Thus, we tested Scenario Generator (SG), a DES software customized for strategic decision planning, in order to show how it could have been used to test general guidelines regarding breast screening program extensions. It is important to say that SG software was implemented in order to support long term strategic and tactical planning evaluation by Public Health managers. Because of this the implementation of new models and clinical pathways had to be very simple in order to ease the utilization to non simulation professionals. SG modeling is then very simple (lacking then of a detailed system modeling), it does not support the definition of daily resource schedules and resource capacity is maily described by number of Full Time Equivalent (FTE) and minimum, average and maximum number of activities that each FTE can perform in a week. In our case it is then impossible to model the fact that some days of the week are dedicated to first level examinations and some others to in-depth invasive and non-invasive ones. Another modeling constraint is the lack of single entities labeling and management, as a consequence no distinction can be made between 2009 and 2010 screened women. Due to those restrictions we decided to test how SG can be used in order to provide a high level information to regional decision planners.

We tested how the system would have behaved in 2010 if the theoretical number of Radiology senology specialists working for the regional health Authority in 2009 had not been changed. Because of SG restrictions we split physicians capacity in two components, first level examination test and second level non-invasive and invasive examinations (see Table 7).

Table 7: First and Second Level Activities per FTE

Activity type	Activities per week per FTE
First level examination test	98
Second level examination	4

As a result we measured a stronger impact of program extension in terms of lead time worsening because the proposed setting would have led to a 45% of first level negative results sent within 15 days (see Figure 4) and 64% within 21 days.

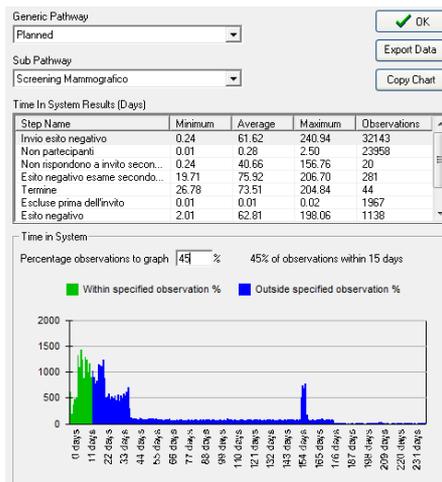


Figure 4: SG 2009-2010 Simulation Output

A resource resizing up to 142 test per 9 Radiology senology specialists would have been necessary in order to increase the performances up to 95% of first level negative results sent within 15 days and 97% within 21 days on average.

In Table 8 we summarize the main differences of the two proposed simulation models and their impact on performance indicators.

Table 8: Comparison of Simulation Models Features

Simul8 operational model	SG tactical model
LOC daily opening hours for screening mammograms	LOC overall weekly capacity for screening mammograms
RSS individual representation and related daily shift	RSSs FTE representation and related weekly workload capacity
Definition of specific days of the week dedicated by RSSs to screening activities (e.g. Monday, Tuesday and Thursday for first level activities)	Definition of the portion of the weekly capacity dedicated by RSSs to predefined activities (e.g. three days for first level activities)
First and second level KPI can be detailed on an yearly, monthly, weekly and daily basis	First and second level KPI can be recorded as cumulative values over the simulation planning horizon

RESULT INTERPRETATION AND CONCLUSION

We analyzed the screening program with two DES software packages in order to show how different tools can help stakeholders that operate at different decision levels. We applied Scenario Generator, an ad-hoc DES software for tactical planning for health systems, in order to develop an high level model that can be used to support

in a quantitative way the definition of regional guidelines. We implemented a detailed DES model using Simul8, a general purpose software, in order to show to local health authority managers how a more detailed model can be used to understand the reasons of long lead times.

The Scenario Generator software acquired by the Regional Agency for health and Social Care of Emilia-Romagna can be used just to provide long term recommendations and to test guidelines implementation. Such recommendations will build upon available regional and local data and will allow the discussion around new services to be provided by the local trusts. A typical questions that could be answered by SG is the annual number of first level analysis that each radiologist should ensure in order to achieve organizational performance goal. Once defined general guidelines it is up to the Local Health Authority management to define if the number of first level test examinations is sustainable by the number of resources available.

It is clear that Scenario Generator software can not provide a detailed system forecast to Local Health Authority planners. The absence of territorial distinctions for mammography machines as well as the impossibility to describe radiologist weekly activities in a detailed way reveals the unfitness of SG as a tool to support operational planning. In order to better control activities in a weekly or annual time horizon is therefore advisable to use a tool like Simul8.

In the proposed application we defined and tested several what-if scenarios and we measured their impact on measured screening lead times. We identified that resource resizing would have been strongly affected more by non-screening activities related to Radiology senology specialists than by population rates of attendance to screening programs. We also evaluated how holiday policies could strongly affect the perception of shortages or surpluses in terms of available resources. That has been proven to be an interesting approach by health managers since during the breast screening planning extension it was not clear if and how much the increased number of Radiology senology specialists would have been able to meet organizational goals. These results helped the Agenzia to asses the need of Decision Support Systems in general, and of operational planning Discrete Event Simulation in particular.

Beemsterboer P.; De Koning, H.; Warmerdam, P.; Boer, R.; Swart, E; Dierks, M.; and Robra, B. 1994. "Prediction of the effects and costs of breast-cancer screening in Germany". *International Journal of Cancer*, 58(5):623-628.

Brailsford, S. and B. Schmidt. 2003. "Towards incorporating human behaviour in models of health care systems: An approach using discrete event

simulation". *European Journal of Operational Research*, 150(1):19 – 31.

Brown, M and L. Fintor. 1993. "Cost effectiveness of breast cancer screening: Preliminary results of a systematic review of the literature". *Breast Cancer Research and Treatment*, 25(2):113–118.

Carter, R.; P. Glasziou; G. van Oortmarssen; H. de Koning; C. Stevenson; G. Salkeld and R. Boer. 1993. "Cost-effectiveness of mammographic screening in Australia". *Australian Journal of Public Health*, 17(1):42–50;

Concannon, K.; M. Elder; K. Hindle; J. Tremble and S. Tse. 2007. "Simulation Modeling with Simul8". <http://www.simtech.hu/dl.php?id=6235>;

Fone D.; S. Hollinghurst; M. Temple; A. Round; N. Lester; A. Weightman; K. Roberts; E. Coyle; G. Bevan and S. Palmer. 2003.. "Systematic review of the use and value of computer simulation modelling in population health and health care delivery". *Journal of Public Health*, 25(4):325–335.

Fryback D.; N.K. Stout; M. A. Rosenberg; A. Trentham-Dietz; V. Kuruchittham and P. L. Remington. 2006. "Chapter 7: The wisconsin breast cancer epidemiology simulation model". *JNCI Monographs*, 2006(36):37–47.

Hunter, D. J.W.; S. M. Drake, S. E.D. Shortt; J. L. Dorland, and N. Tran. 2004. "Simulation modeling of change to breast cancer detection ageeligibility recommendations in ontario,2002-2021". *Cancer Detection and Prevention*,28(6):453 – 460.

Michaelson, J. S.; E. Halpern, and D. B. Kopans. 1999. "Breast cancer: Computer simulation method for estimating optimal intervals for screening". *Radiology*, 212(2):551–560.

Naldoni, C.; A. Finarelli and R. Mignani. 2012. "Protocollo diagnostico terapeutico dello screening per la diagnosi precoce dei tumori della mammella della regione emiliaromagna". *Contributi Servizio Sanità Pubblica Regione Emilia-Romagna*, (69).

Paci, E.; R. Boer; M. Zappa; H. J. de Koning; G. J. van Oortmarssen; E. Crocetti; D. Giorgi; M. Rosselli del Turco; and J. D. Habbema. 1995. "A model-based prediction of the impact on reduction in mortality by a breast cancer screening programme in the city of Florence, Italy". *Eur. J. Cancer*, 31A(3):348–353.

ANDREA LODI received the PhD in System Engineering from the University of Bologna in 2000 and he has been Herman Goldstine Fellow at the IBM Mathematical Sciences Department, NY in 2005-2006.

He is full professor of Operations Research at DEI, University of Bologna since 2007. He is author of more than 70 publications in the top journals of the field of Operations Research; he serves as Associated Editor for several prestigious journals in the area. Andrea Lodi is currently network coordinator of the EU project FP7-PEOPLE-2012-ITN and ICT COST ACTION TD1207, and, since 2006, consultant of the IBM CPLEX research and development team.

PAOLO TUBERTINI studied Industrial and Management Engineering at the University of Bologna and obtained his degree in 2010. In 2014 he received the PhD in Automatic Control and Operational Research at the "Dipartimento di Ingegneria dell'Energia Elettrica e dell'Informazione "Guglielmo Marconi" – DEI" University of Bologna. During the PhD research period he has collaborated with the Regional Health Authority of Emilia-Romagna Region to support Operational Research applications to regional health organizations.

ROBERTO GRILLI is an epidemiologist and health services researcher, since 2006 he is the Director of the Regional Agency for Health and Social Care of Emilia-Romagna, the organization supporting the Regional Health Authority of Emilia-Romagna. He has been working extensively in the area of health technology assessment, quality assessment of the scientific literature, quality of care evaluation, development and implementation of practice guidelines, governance in healthcare organizations. He has been involved in several national and international projects, being also (from 1994 up to 2006) among the members of the editorial team of the Cochrane Effective Practice and Organization of Care (EPOC) Review Group. He is author and co-author of 186 scientific publications, and of three books.

FRANCESCA SENESE holds a Master in Public Health (EUROPUBHEALTH) and since 2010 she collaborates with the Regional Agency for Health and Social Care of Emilia-Romagna. Before joining the Regional Agency she worked at WHO Regional Office for Europe within the Health Intelligence Unit (HIU). Her interests are in the field of health economics and health services research. As part of her current activity she collaborates with the DEI team to implement Operational Research techniques to the regional health system.

AUTHOR INDEX

- 57 Aarseth, Johnny
603 Abaev, Pavel O.
676 Abaide, Alzenira
382 Adamek, Milan
319 Adamu, Abdullahi S.
57, 87 AEsoy, Vilmar
115
651 Aikala, Antti
81, 643 Alaliyat, Saleh
501 Allidina, Alnoor
333 Alvarez, Tamara
359 Anyameluhor, Nnamdi
691 Archetti, Claudia
18 Ayed, Hedi
431 Bacchetti, Andrea
775 Balke, Tina
475 Baranowski, Bartosz
620 Barbey, Hans-Peter
226, 319 Bargiela, Andrzej
364
136 Baronio, Fabio
709 Barra, Carlos
226 Baskaran, Geetha
147 Bau, Marco
326, 633 Belyakov, Stanislav L.
326 Belykova, Marina L.
577, 590 Bening, Vladimir
596
179, 421 Berman, Sigal
676 Bernardon, Daniel Pinheiro
633 Bershtein, Leonid S.
448 Bertazzi, Luca
501 Blaszczyk, Jacek
247, 254 Bobal, Vladimir
261, 267
159 Bobasu, Eugen
41 Bonomi, Germano
523 Boryczko, Krzysztof
11, 32 Bossomaier, Terry
326, 633 Bozhenyuk, Alexander V.
718 Bravo, Giangiacomo
670 Breidbach, Joerg
142, 153 Brojboiu, Maria
551 Bruneel, Herwig
57, 87 Bunes, Oyvind
333 Burguillo, Juan C.
682 Bye, Robin T.
333, 495 Byrski, Aleksander
725 Cabota, Juan Bautista
709 Canessa, Enrique
115 Cao, Yanran
558 Caraccio, Ilaria
789 Carnevale, Claudio
480 Cerotti, Davide
454 Chabrol, Michelle
709 Chaigneau, Sergio E.
782 Chappin, Emile
319 Chong, Siang-Yew
57, 87 Chu, Yingguang
25 Cicirelli, Franco
551 Claeys, Dieter
185 Cleophas, Catherine
136 Conforti, Matteo
333 Covelo, Jose
613 D'Apice, Ciro
670 Danner, Christian
371, 393 Davendra, Donald
171 De Angelis, Costantino
613 De Nicola, Carmine
71 Dibbern, Christoph
732 Dilaver, Oezge
530 Dobre, Ciprian
752 Doemoetoer, Barbara
41 Donzella, Antonietta
247, 254 Dostal, Petr
273, 297
11, 32 Duncan, Roderick
625 Eisler, Cheryl
523 Faber, Lukasz
147 Farran, Mohamad
147 Ferrari, Marco
147 Ferrari, Vittorio
179 Fink, Lior
11 Finlayson, Max C.
789 Finzi, Giovanna
676 Fonini, Julio
401, 411 Forsman, Pekka
185 Fuerstenau, Daniel
206 Furian, Nikolaus
475 Gaciarz, Tomasz
603 Gaidamaka, Yuliya V.
676 Garcia, Vinicius Jacques
18 Gateau, Benjamin
285 Gazdos, Frantisek
47 Geuter, Juergen
570, 596 Gorshenin, Andrey

480 Gribaudo, Marco
 583, 596 Grigoryeva, Maria E.
 796 Grilli, Roberto
 725 Grimaldo, Francisco
 6 Grimm, Volker
 214 Grossmann, Tobias
 565 Grusho, Nick A.
 565 Grusho, Alexander A.
 515 Grzonka, Daniel
 64 Gubian, Paolo
 47, 71 Hahn, Axel
 401, 411 Halme, Aarne
 745 Haskell, Evan C.
 101 Hatledal, Lars I.
 657 Herrmann, Frank
 421 Hershkovitz Cohen, Anat
 94 Hildre, Hans Petter
 670 Husslein, Thomas
 480 Iacono, Mauro
 279 Iliev, Vasil
 309 Incerti, Giovanni
 487 Ireno, Dawid R.
 153, 159 Ivanov, Sergiu
 142, 153 Ivanov, Virginia
 377 Jasek, Roman
 782 Jensen, Thorben
 775 Johnson, Peter George
 758 Juhasz, Peter
 279 Kaneva, Maria
 461 Kersten, Wolfgang
 131 Ketnere, Elena
 18 Khadraoui, Djamel
 664 Kim, Byeong Soo
 664 Kim, Tag Gon
 495, 523 Kisiel-Dorohinicki, Marek
 515 Kolodziej, Joanna
 387 Kominek, Ales
 346, 382 Kominkova Oplatkova,
 387, 393 Zuzana
 279 Koprinkova-Hristova, Petia
 570, 577 Korolev, Victor
 583, 590
 596
 279 Kostov, Georgi
 775 Kotthoff, Lars
 340, 346 Kotyrba, Martin
 267, 273 Krhovjak, Adam
 254, 261 Kubalcik, Marek
 570 Kuzmin, Victor
 292 La Vecchia, Giovina Marina
 382 Lapkova, Dora
 670 Lauf, Wolfgang
 664 Lee, Sun Ju
 541 Levis, Alexander H.
 94 Li, Wei
 57 Lien, Alf Helge
 193, 235 Liguori, Arturo
 171 Locatelli, Andrea
 796 Lodi, Andrea
 765 Lotzmann, Ulf
 333 Loureiro, Miguel
 551 Maertens, Tom
 570 Malakhov, Dmitry
 501 Malinowski, Krzysztof
 613 Manzo, Rosanna
 285 Marholt, Jiri
 508 Marks, Michal
 297 Maslan, Martin
 319, 364 Maul, Tomas H.
 166 Melnik, Roderick
 131 Mesnajevs, Aleksandrs
 147 Modotto, Daniele
 639 Mueller, Christian
 701 Mueller, Georg P.
 352 Nahodil, Pavel
 279 Naydenova, Vessela
 64 Neri, Bruno
 64 Neri, Paolo
 765 Neumann, Martin
 508 Niewiadomska-Szynkiewicz,
 Ewa
 25 Nigro, Libero
 206 O'Sullivan, Michael
 448 Ohlmann, Jeffrey
 7 Onorato, Miguel
 81 Osen, Ottar L.
 454 Paris, Jean-Luc
 558, 603 Pechinkin, Alexander V.
 789 Pederzoli, Anna
 333 Peleteiro, Ana
 11, 32 Perez-Mujica, Luisa
 202, 214 Petermann, Arne
 309 Petrogalli, Candida
 359 Peytchev, Evtim
 480 Piazzolla, Pietro
 64 Piccinelli, Mario
 122 Piros, Attila
 475 Plichta, Anna
 371, 382 Pluhacek, Michal
 393

292, 304 Pola, Annalisa
 530 Pop, Florin
 159 Popescu, Dan
 279 Popova, Silviya
 166 Prabhakar, Sanjay
 25 Pupo, Francesco
 226 Qu, Rong
 709 Quezada, Ariel
 179 Raphaeli, Orit
 159 Rasvan, Vladimir
 558, 603 Razumchik, Rostislav V.
 108 Rekdalsbakken, Webjorn
 551 Reveil, Bert
 738 Righi, Simone
 193, 235 Romanin-Jacur, Giorgio
 179 Rosenfeld, Liron
 454 Royer, Johan
 326, 633 Rozenberg, Igor N.
 651 Ruuska, Pekka
 411 Saarinen, Jari
 461 Saeed, Muhammad Amad
 603 Samouylov, Konstantin E.
 81, 101 Sanfilippo, Filippo
 108, 643
 577 Satin, Yakov
 142 Savescu, Andrei
 590 Savushkin, Vladislav A.
 682 Schaathun, Hans G.
 185 Schinzel, Johannes
 71 Schweigert, Soeren
 166 Sebetci, Ali
 670 Seidl, Markus
 796 Senese, Francesca
 371, 377 Senkerik, Roman
 387, 393
 577 Shilova, Galina
 570, 577 Shorgin, Sergey Ya.
 590 Shunkov, Egor I.
 202 Simon, Alexander
 570 Skvortsova, Nina
 309 Solazzi, Luigi
 292, 304 Soltani, Mahdi
 664 Song, Hae Sang
 691 Speranza, M. Grazia
 725 Squazzoni, Flaminio
 115 Stene, Anne
 159 Stinga, Florin
 530 Stoica, Cosmin-Stefan
 41 Subieta, M.
 377 Svejda, Jaromir
 758 Szaz, Janos
 475 Szominski, Szymon
 738 Takacs, Karoly
 254, 261 Talas, Stanislav
 267, 273
 515 Tao, Jie
 5 Thurner, Stefan
 565 Timonina, Elena E.
 441 Trunfio, Roberto
 796 Tubertini, Paolo
 364 Turcsany, Diana
 789 Turrini, Enrico
 752 Varadi, Kata
 122 Varga, Eszter
 122 Vidovics, Balazs
 758 Vidovics-Dancs, Agnes
 352 Vitku, Jaroslav
 206 Voessner, Siegfried
 247, 297 Vojtesek, Jiri
 340, 346 Volna, Eva
 789 Volta, Marialuisa
 206 Walker, Cameron
 551 Walraevens, Joris
 651 Weiss, Robert
 625 Wesolkowski, Slawomir
 625 Wojtaszek, Daniel T.
 304 Xu, Qiang
 401, 411 Ylikorpi, Tomi
 643 Yndestad, Harald
 541 Yousefi, Bahram
 377 Zak, Roman
 431 Zanardini, Massimo
 577, 583 Zeifman, Alexander
 590, 596
 371, 393 Zelinka, Ivan
 41 Zenoni, Aldo
 87, 94 Zhang, Houxiang
 101

