

Copyright

© ECMS2016

Printed: ISBN: 978-0-9932440-2-5

**European Council for Modelling
and Simulation**

CD: ISBN: 978-0-9932440-3-2

Cover pictures

© front side:

Peter Ferstl, Stadt Regensburg

©back side: OTH Regensburg

Printed by

Digitaldruck Pirrot GmbH

66125 Sbr.-Dudweiler, Germany

PROCEEDINGS

30th European Conference on Modelling and Simulation ECMS 2016

May 31st – June 3rd, 2016
Regensburg, Germany

Edited by:

Thorsten Claus

Frank Herrmann

Michael Manitz

Oliver Rose

Organized by:

ECMS - European Council for Modelling and Simulation

Hosted by:

OTH – Ostbayerische Technische Hochschule Regensburg,
Germany

Sponsored by:

OTH – Ostbayerische Technische Hochschule Regensburg,
Germany

CC-SE Competence Center for Software Engineering,
Regensburg, Germany

OptWare GmbH, Regensburg, Germany

SimPlan AG, Germany

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 2016 ORGANIZATION

Conference Chair

Frank Herrmann

OTH Regensburg
Germany

Conference Co-Chair

Thorsten Claus

Technical University Dresden
Germany

Programme Chair

Michael Manitz

University of Duisburg-Essen
Germany

Programme Co-Chair

Oliver Rose

University of the Federal Armed Forces Munich
Germany

Assistant Conference Chair

Maximilian Munniger

OTH Regensburg, University of Duisburg-Essen
Germany

President of European Council for Modelling and Simulation

Khalid Al-Begain

University of South Wales
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

Effectively Operating Simulation

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

Co-Chair: **Arne Petermann**
Free University of Berlin, Germany
Director, Institute for Quality and Management
in the Health Care System International Academy (INA)

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 and Economics and Social Science

Track Chair: **Kata Váradi**
Corvinus University of Budapest, Hungary

Co-Chairs:
Barbara Dömötör
Corvinus University of Budapest, Hungary

Ágnes Vidovics-Dancs
Corvinus University of Budapest, Hungary

Applied Modelling and Simulation

Track Chair: **Gaby Neumann**
Technical University of Applied Sciences Wildau, Germany

Co-Chair: **Edward J. Williams**
University of Michigan-Dearborn, USA

Modelling, Simulation and Control of Technological Processes

Track Chair: **Jiří Vojtěšek**

Tomas Bata University in Zlín, Czech Republic

Co-Chairs:

Petr Dostál

Tomas Bata University in Zlín, Czech Republic

František Gazdoš

Tomas Bata University in Zlín, Czech Republic

Simulators for Virtual Prototyping and Training

Track Chair: **Ottar L. Osen**

NTNU in Ålesund, Norway

Co-Chairs:

Henrique Murilo Gaspar

NTNU in Ålesund, Norway

Robin Trulssen Bye

NTNU in Ålesund, Norway

Simulation and Optimization

Track Chair: **Frank Herrmann**

OTH Regensburg, Germany

Co-Chairs:

Thorsten Claus

Technical University Dresden, Germany

Michael Manitz

University of Duisburg-Essen, Germany

High Performance Modelling and Simulation

Track Chair: **Mauro Iacono**

Seconda Università degli Studi di Napoli, Italy

Co-Chairs:

Daniel Grzonka

Cracow University of Technology, Poland

Rostislav V. Razumchik

Institute of Informatics Problems FRC CSC RAS,
Russia

Honorary Track Chair:

Joanna Kolodziej

Cracow University of Technology, Poland

IPC Members in Alphabetical Order

Pavel Abaev, Peoples' Friendship University of Russia, Russia

Frederic Amblard, Université Toulouse 1 Capitole, 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

Enrico Barbierato, Seconda Università degli Studi di Napoli, Italy

Simona Bernardi, Centro Universitario de la Defensa, Spain

Nik Bessis, Edgehill University, United Kingdom

Håkon Bjørlykke, Rolls-Royce Marine, Norway

Frøy Birte Bjørneseth, Rolls-Royce Marine, Norway

Vladimir Bobal, Tomas Bata University in Zlin, Czech Republic

Riccardo Boero, Los Alamos National Laboratory, USA

Aleksander Byrski, AGH University of Science and Technology, Poland

Krzysztof Cetnarowicz, AGH University of Science and Technology, Poland

Petr Chalupa, Tomas Bata University in Zlin, Czech Republic

Marina Chukalina, Russian Academy of Science, Russia

Franco Cicirelli, University of Calabria, Italy

Catherine Cleophas, RWTH Aachen, Germany

Valentin Cristea, University Politehnica of Bucharest, Romania

Donald Davendra, Central Washington University, USA

Antinisca Di Marco, Università degli Studi dell'Aquila, Italy

Ciprian Dobre, University Politehnica of Bucharest, Romania

František Dušek, University of Pardubice, Czech Republic

Tobias Ebbing, Berlin, Germany

Massimo Ficco, Seconda Università degli Studi di Napoli, Italy

Daniel Fürstenau, Free University Berlin, Germany

Charlotte Gerritsen, VU University of Amsterdam, The Netherlands

Amineh Ghorbani, TU Delft, The Netherlands

Horacio Gonzalez-Velez, National College of Ireland, Ireland

Marco Gribaudo, Politecnico di Milano, Italy

Dag Sverre Grønmyr, Rolls-Royce Marine, Norway

Alexander Grusho, Institute of Informatics Problems, FRC CSC RAS, Russia

Ibrahim A. Hameed, NTNU in Ålesund, Norway

Dániel Havran, Corvinus University of Budapest, Hungary
Daniel Honc, University of Pardubice, Czech Republic
Mark Hoogendorn, VU University of Amsterdam, The Netherlands
Shams Mahmood Imam, Rice University, USA
Teruaki Ito, The University of Tokushima, Japan
Agnieszka Jakóbbik, Cracow University of Technology, Poland
Michal Janosek, University of Ostrava, Czech Republic
Jácint Juhász, Babes-Bolyai University, Hungary
Bogumil Kaminski, Warsaw School of Economics, Poland
Petia Koprinkova, Bulgarian Academy of Sciences, Bulgaria
Victor Korolev, Moscow State University, Russia
Igor Kotenko, SPIIRAS, Russia
Martin Kotyrba, University of Ostrava, Czech Republic
Imola Kovács, Babes-Bolyai University, Romania
Marek Kubalcik, Tomas Bata University in Zlin, Czech Republic
Jane Labadin, University of Malaysia Sawarak, Malaysia
Alexander H. Levis, George Mason University, USA
Guoyuan Li, NTNU in Ålesund, Norway
Catalina Lladò, Universitat de le Illes Balears, Spain
Christian Lutz, University Duisburg-Essen, Germany
Fernando Luz, Numerical Offshore Tank, Univ. of Sao Paulo, Norway, Brazil
Andrea Marin, Università "Ca' Foscari" di Venezia, Italy
Michal Marks, Research and Academic Computer Network, Poland
Stefano Marrone, Seconda Università degli Studi di Napoli, Italy
Agnieszka Mars, Jagiellonian University, Poland
Katarina Matějčková, VUCHT, Slovakia
Radek Matusu, Tomas Bata University in Zlin, Czech Republic
Nicolas Meseth, Deloitte Consulting, Germany
Marcio Michiharu Tsukamoto, Num. Offshore Tank, Univ. of Sao Paulo, Norway, Brazil
Christian Müller, University of Applied Sciences in Wildau, Germany
Maximilian Munniger, University of Duisburg-Essen, Germany
Pavel Nahodil, Czech Technical University in Prague, Czech Republic
Valeriy Naumov, Service Innovation Research Institute, Finland
Catalin Negru, University Politehnica of Bucharest, Romania

Ewa Niewiadomska-Szynkiewicz, Warsaw University of Technology and NASK, Poland
Libero Nigro, University of Calabria, Italy
Dmitry P. Nikolaev, Russian Academy of Science, Russia
Lars Nolle, Jade University of Applied Science, Germany
Jakub Novak, Tomas Bata University in Zlin, Czech Republic
Francesco Palmieri, Università degli Studi di Salerno, Italy
Falk Pappert, University of the Bundeswehr Munich, Germany
Libor Pekar, Tomas Bata University in Zlin, Czech Republic
Michal Pluhacek, Tomas Bata University in Zlin, Czech Republic
Florin Pop, University Politehnica of Bucharest, Romania
Michael Puhle, Passauer University, Germany
Francesco Pupo, University of Calabria, Italy
Simone Righi, Hungarian Academy of Sciences, Hungary
Boris Rohal-Ilkiv, Slovak University of Technology in Bratislava, Slovakia
Michael Römer, Martin-Luther-Universität Halle-Wittenberg, Germany
Konstantin Samouylov, Peoples' Friendship University, Russia
Thomas Schulze, University of Magdeburg, Germany
Maximilian Selmair, Technical University Dresden, Germany
Leonid Sevastyanov, Peoples' Friendship University, Russia
Oleg Shestakov, Moscow State University, Russia
Sergey Ya. Shorgin, Institute of Informatics Problems, FRC CSC RAS, Russia
Markus Siegle, University of the German Armed Forces, Germany
Anders Skoogh, Chalmers University of Technology, Sweden
Mojca Indihar Štemberger, University of Ljubljana, Slovenia
Girts Strazdins, NTNU in Ålesund, Norway
Grażyna Suchacka, Opole University, Poland
János Száz, Corvinus University of Budapest, Hungary
Magdalena Szmajdych, Cracow University of Technology, Poland
Julia Szota-Pachowicz, Jagiellonian University, Poland
Armando Tacchella, Università degli Studi di Genova, Italy
Enrico Teich, Technical University Dresden, Germany
Pietro Terna, University of Torino, Italy
Klaus Troitzsch, University of Koblenz-Landau, Germany
Christopher Tubb, University of South Wales, United Kingdom

Tobias Uhlig, University of the Federal Armed Forces Munich, Germany
Nikolai Ushakov, Norwegian University of Science and Technology, Norway
Enrico Vicario, Università degli Studi di Firenze, Italy
Jaroslav Vitku, Technical University in Prague, Czech Republic
Rune Volden, Ulstein Power & Control AS, Norway
Eva Volna, University of Ostrava, Czech Republic
Andrzej Wilczyński, Cracow University of Technology, Poland
Victor Zakharov, Institute of Informatics Problems, FRC CSC RAS, Russia
Armin Zimmermann, Technical University of Ilmenau, Germany

PREFACE

“There is something irreversible about acquiring knowledge; and the simulation of the search for it differs in a most profound way from the reality.”

by J. Robert Oppenheimer – American Theoretical Physicist (1904-1967)
in: “Physics in the Contemporary World” lecture at M.I.T. (25th November 1947)

For many years there is a stable growth of simulation in research with important applications. Over 30 years the European Conference on Modelling and Simulation reflects this development in many fields. Thus, this proceedings of the 30th European Conference on Modelling and Simulation (ECMS) from May 31st until June 3rd, 2016, hosted by OTH Regensburg in Germany, represents the current status.

The human capacity to abstract complex systems and phenomena into simplified models has played a critical role in the rapid evolution of our modern industrial processes and scientific research. As a science and an art, modelling and simulation have been one of the core enablers of this remarkable human trace, and have become a topic of great importance for researchers and practitioners. The increasing availability of massive computational resources and interconnectivity has helped tremendous advances in the field, collapsing previous barriers and redefining new horizons for its capabilities and applications. The scope of the accepted papers shows that this ECMS conference brings together researchers, engineers, applied mathematicians and practitioners working on these topics. In detail, these papers include methodologies, technologies, applications and tools for modelling and simulation. In our thinking, also caused by the review through recognized experts, the accepted papers demonstrate new and innovative solutions. They also highlight technical issues and challenges in this field.

The high quality of the ECMS 2016 program is enhanced by the two keynote lectures, delivered by distinguished speakers who are renowned experts in their fields: Professor Dr. Andrea Matta (Department of Industrial Engineering and Management at Shanghai JiaoTong University) and Dr. Thomas Hußlein (Managing Director of OptWare GmbH in Regensburg).

The timetable leaves enough room for various discussions and socializing. Especially, the comprehensive program with the conference dinner during a boat trip on the Danube serves to build and enhance social contacts amongst the participants.

Building an interesting and successful programme for the conference required the dedicated effort of many people; we thank them all. Especially, the authors, whose research and development efforts are recorded here, deserve our thanks. We thank the members of the Programme Committee, the Local Organization Committee and additional reviewers for their diligence and expert reviewing. Last but not least, we thank the invited distinguished keynote speakers for their invaluable contribution and for taking the time to synthesize and deliver their talks. The generosity of all sponsors allows some highlights in the social program.

All the contributions in this conference may stimulate further research. In addition, we hope that you will get valuable insights in your research area, and you will have inspiring discussions. Please enjoy the ECMS 2016 programme and your stay in Regensburg, Germany.

Thorsten Claus, Frank Herrmann, Michael Manitz and Oliver Rose
May 2016

TABLE OF CONTENTS

Plenary Talks - Abstracts

Discrete Event Optimization: Theory, Applications And Future Challenges <i>Andrea Matta</i>	5
Data Analytics, Model Generation And Optimization Algorithms - A Perfect Match? <i>Thomas Husslein</i>	7

Agent-Based Simulation

A Learning Agent For A Multi-Agent System For Project Scheduling In Construction <i>Florian Wenzler, Willibald A. Guenther</i>	11
Agent-Based Model Continuity Of Stochastic Time Petri Nets <i>Franco Cicirelli, Libero Nigro, Paolo F. Sciammarella</i>	18
Frontier Based Multi Robot Area Exploration Using Prioritized Routing <i>Rahul Sharma K., Daniel Honc, Frantisek Dusek, Gireesh Kumar T.</i>	25

Effectively Operating Simulation

Simulation-Based Performance Measurement: Assessing The Purchasing Process In A Public University <i>Pasquale Legato, Lidia Malizia, Rina Mary Mazza</i>	33
Path Dependence In Hierarchical Organizations: The Influence Of Environmental Dynamics <i>Arne Petermann, Alexander Simon</i>	41

Applied Modelling and Simulation

Reachability Of Fractional Continuous-Time Linear Systems Using The Caputo-Fabrizio Derivative

Tadeusz Kaczorek53

Simulation Improves Operations At A Specialized Takeout Restaurant

*Sapthagirishwaran Thennal Sivaramakrishnan,
Shanmugasundaram Chandrasekaran, Jennifer Dhanapal,
Paul Ajaydivyan Jeya Sekar, Edward J. Williams*59

Making Of Credible Permeability Maps For Layers Of Hydrogeological Model Of Latvia

Aivars Spalvins, Inta Lace, Kaspars Krauklis66

Concept Hierarchies For Sensor Data Fusion In The Cognitive IoT

Franco Cicirelli, Giandomenico Spezzano73

A Simulation Based Study Of The Effect Of Truck Arrival Patterns On Truck Turn Time In Container Terminals

Ahmed E. Azab, Amr B. Eltawil80

3D Simulation Modeling Of Yard Operation In A Container Terminal

Jingjing Yu, Chen Liang, Guolei Tang87

Generic Reaction-Diffusion Model For Transmission Of Mosquito-Borne Diseases: Results Of Simulation With Actual Cases

Cynthia Mui Lian Kon, Jane Labadin93

Some GPSS Opportunities For Modeling Of Timestamp Ordering In DDBMS And Simulation Investigations

Svetlana Vasileva100

An Assessment Of Pharmacological Properties Of *Schinus* Essential Oils - A Soft Computing Approach

*Jose Neves, M. Rosario Martins, Fatima Candeias, Silvia Arantes,
Ana Piteira, Henrique Vicente*107

Truck Arrival Management At Maritime Container Terminals

Daniela Ambrosino, Lorenzo Peirano114

An Application Of Discrete Event Simulation On Order Picking Strategies: A Case Study Of Footwear Warehouses

Thananya Wasusri, Prasit Theerawongsathon121

Finance and Economics and Social Science

Modeling Inferential Minds In Conceptual Space

Carlos Barra, Enrique Canessa, Sergio E. Chaigneau 131

Individual-Level Simulation Model For Cost Benefit Analysis In Healthcare

Nagesh Shukla, Vu Lam Cao, Van Hoang Phuong, Marian Shanahan, Allison Ritter, Pascal Perez..... 138

Optimal Fiscal Policies After The “Great Recession”: A Case Study For Slovenia

Reinhard Neck, Dmitri Blueschke, Klaus Weyerstrass..... 145

Friendship Of Stock Indices

Laszlo Nagy, Mihaly Ormos 152

The Use Of Cluster Analysis For Demographic Policy Development: Evidence From Russia

Oksana Shubat, Anna Bagirova, Abilova Makhabat, Anton Ivlev..... 159

Classical And Novel Risk Measures For A Stock Index On A Developed Market

Julia Timea Nagy, Balint Zsolt Nagy, Jacint Juhasz 166

Taxation And Corporate Performance: Less Is More

Peter Juhasz, Kata Varadi 172

Mental Framing In Risk-Aversion Dynamics An Empirical Investigation Of Intertemporal Choice

Mihaly Ormos, Dusan Timotity..... 179

Intertemporal Choice And Dynamics Of Risk Aversion

Mihaly Ormos, Dusan Timotity..... 185

Estimation Of Customer Default Based On Behavioural Variables

Nora Felfoeldi-Szuecs..... 192

Factors Affecting Household Participation In Solid Waste Management Segregation And Recycling In Bangkok, Thailand

Atthirawong Walailak 198

A Plea For Microsimulation

Marc Hannappel 204

**Small Shipment Delivery's Quality Improvement In Cities
With Unstable Traffic**

Pavel Patlins, Remigijs Pocs211

Tracking Business Trends – Dilemmas Of Measurement

Peter Juhasz, Janos Szaz, Kata Varadi, Agnes Vidovics-Dancs217

Simulation of Intelligent Systems

Fast 3D Hough Transform Computation

*Egor. I. Ershov, Arseniy P. Terekhin, Simon M. Karpenko,
Dmitry P. Nikolaev, Vassili V. Postnikov*227

New Approach Of Constant Resolving Of Analytical Programming

Tomas Urbanek, Zdenka Prokopova, Radek Silhavy, Ales Kuncar231

Analytical Programming With Extended Individuals

*Adam Viktorin, Michal Pluhacek, Zuzana Kominkova Oplatkova,
Roman Senkerik*237

**Multi-Chaotic Differential Evolution For Vehicle Routing Problem
With Profits**

Adam Viktorin, Dusan Hrabec, Michal Pluhacek.....245

Study On Swarm Dynamics Converted Into Complex Network

*Michal Pluhacek, Roman Senkerik, Jakub Janostik,
Adam Viktorin, Ivan Zelinka*252

**On The Simulation Of Complex Chaotic Dynamics
For Chaos Based Optimization**

*Roman Senkerik, Michal Pluhacek, Adam Viktorin,
Zuzana Kominkova Oplatkova*258

**Simulation Of Submarine Groundwater Discharge Of Dissolved Organic
Matter Using Cellular Automata**

Lars Nolle, Holger Thormaehlen, Harald Musa.....265

Design And Simulation Of Integrated EMI Filter

Jens Werner, Jennifer Schuett, Guido Notermans.....270

Modelling, Simulation and Control of Technological Processes

Identification And LQ Digital Control Of A Set Of Equal Cylinder Atmospheric Tanks – Simulation Study <i>Vladimir Bobal, Petr Dostal, Marek Kubalcik, Stanislav Talas</i>	279
Discrete Method For Estimation Of Time-Delay Outside Of Sampling Period <i>Stanislav Talas, Vladimir Bobal, Adam Krhovjak, Lukas Rusar</i>	287
Nonlinear Simulink Model Of Magnetic Levitation Laboratory Plant <i>Petr Chalupa, Martin Maly, Jakub Novak</i>	293
Linear Predictive Control Of Nonlinear Time-Delayed Model Of Liquid-Liquid Stirred Heat Exchanger <i>Radek Holis, Vladimir Bobal</i>	300
State-Space Predictive Control Of Two Liquid Tanks System With Constraints Of Process Variables <i>Lukas Rusar, Stanislav Talas, Adam Krhovjak, Vladimir Bobal</i>	307
Cascade Control Of A Tubular Chemical Reactor Using Nonlinear Part Of Primary Controller <i>Petr Dostal, Jiri Vojtesek, Vladimir Bobal</i>	313
Continuous-Time Vs. Discrete-Time Identification Models Used For Adaptive Control Of Nonlinear Process <i>Jiri Vojtesek, Petr Dostal</i>	320
Optimal Gain Scheduled Controller For A Two Funnel Liquid Tanks In Series <i>Adam Krhovjak, Petr Dostal, Stanislav Talas, Lukas Rusar</i>	327
Bond Graph Model Of A Water Heat Exchanger <i>Toufik Bentaleb, Minh Tu Pham, Damien Eberard, Wilfrid Marquis-Favre</i>	333
Modelling A PCT40 Heat Exchanger For Control Purposes <i>Frantisek Gazdos, Daniel Macek</i>	340
Predictive Control Of Three-Tank-System Utilizing Both State-Space And Input-Output Models <i>Marek Kubalcik, Vladimir Bobal</i>	347

Predictive Control Of Differential Drive Mobile Robot Considering Dynamics And Kinematics	
<i>Rahul Sharma K., Daniel Honc, Frantisek Dusek</i>	354
Predictive And Feedback Linearizing Control Of <i>Chlamydomonas Reinhardtii</i> Photoautotrophic Growth Process	
<i>Florin Stinga, Emil Petre</i>	361
Greenhouse Modeling And Simulation Framework For Extracting Optimal Control Parameters	
<i>Byeong Soo Kim, Bong Gu Kang, Tag Gon Kim, Hae Sang Song</i>	368
Dominant Spectrum Assignment For Neutral Time Delay Systems: A Study Case	
<i>Libor Pekar, Roman Prokop</i>	374
Evaluation Of The Primary Metabolism Of Monocultures And Yoghurt Starters With The Participation Of Urease-Deficient <i>Streptococcus Thermophilus</i> Strains	
<i>Ivan Petelkov, Rositsa Denkova, Bogdan Goranov, Vesela Shopska, Georgi Kostov, Zapryana Denkova, Nadya Ninova-Nikolova, Zoltan Urshev, Svetlana Minkova</i>	381
Modeling And Analysis Of Spin Splitting In Strained Graphene Nanoribbons	
<i>Sanjay Prabhakar, Roderick Melnik, Luis Bonilla</i>	388
Using Of Orientation Sensor CHR 6-DM In Security Technologies	
<i>Milan Adamek, Petr Neumann, Martin Pospisilik</i>	393
Schottky Diode Replacement By Transistors: Simulation And Measured Results	
<i>Martin Pospisilik</i>	399

Simulation and Optimization

A New Approach For The Bullwhip Effect

Hans-Peter Barbey407

The Business Process Simulation Standard (BPSIM): Chances And Limits

Ralf Laue, Christian Mueller.....413

Hybrid Model Of Human Mobility For DTN Network Simulation

Alexander Privalov, Alexander Tsarev.....419

Optimization Of A Heat Radiation Intensity And Temperature Field On The Mould Surface

Jaroslav Mlynek, Roman Knobloch, Radek Srb.....425

Social And Ecological Capabilities For A Sustainable Hierarchical Production Planning

Marco Trost, Thorsten Claus, Enrico Teich, Maximilian Selmair, Frank Herrmann.....432

A Supply Chain Optimization Framework For CO₂ Emission Reduction: Case Of The Netherlands

Narayen Kalyanarengan Ravi, Edwin Zondervan, Martin Van Sint Annaland, J.C. (Jan) Fransoo, Johan Grievink.....439

Hybridising Local Search With Branch-And-Bound For Constrained Portfolio Selection Problems

Fang He, Rong Qu.....446

Modelling And Optimization Of The Second-Harmonic Radiation Pattern In Dielectric Nanoantennas

Davide Rocco, Luca Carletti, Andrea Locatelli, Costantino De Angelis, Valerio F. Gili, Giuseppe Leo.....453

Abstraction On Network Model Under Interoperable Simulation Environment

Bong Gu Kang, Byeong Soo Kim, Tag Gon Kim.....460

Models And Algorithms For Abilities Evaluation Of Active Moving Objects Control System

Boris Sokolov, Vladimir Kalinin, Sergey Nemykin, Dmitry Ivanov.....467

Emergency Prediction In Electric Utilities: A Case Study From South Brazil

Iochane Guimaraes, Vinicius Jacques Garcia, Daniel Pinheiro Bernardon, Julio Fonini.....474

Investigation Of Genetic Operators And Priority Heuristics for Simulation Based Optimization Of Multi-Mode Resource Constrained Multi-Project Scheduling Problems (MMRCMPSP)	
<i>Mathias Kuehn, Taiba Zahid, Michael Voelker, Zhugen Zhou, Oliver Rose</i>	481
Job Shop Scheduling With Flexible Energy Prices	
<i>Maximilian Selmair, Thorsten Claus, Marco Trost, Andreas Bley, Frank Herrmann</i>	488
Model-Based Approach To Study Hot Rolling Mills With Data Farming	
<i>Dariusz Krol, Renata Słota, Jacek Kitowski, Lukasz Rauch, Krzysztof Bzowski, Maciej Pietrzyk</i>	495
Future Demand Uncertainty In Personnel Scheduling: Investigating Deterministic Lookahead Policies Using Optimization And Simulation	
<i>Michael Roemer, Taieb Mellouli</i>	502
Optimal Production Volume Of Rubber Gloves Mold For Rubber Gloves Production Planning	
<i>Tuanjai Somboonwiwat, Chorkaew Jaturanonda, Nattapong Chotpan, Kanogkan Leerojanaprapa</i>	508
Optimal Scheduling Of Two-Stage Reentrant Hybrid Flow Shop For Heat Treatment Process	
<i>Noppachai Chalardkid, Tuanjai Somboonwiwat, Chareonchai Khompatraporn</i>	515
Mathematically Modelling HCG In Women With Gestational Trophoblastic Disease Using Exponential Interpolation	
<i>Catherine Costigan, Sabin Tabirca, John Coulter, Ernest Scheiber</i>	520

Simulators for Virtual Prototyping and Training

vMannequin: A Fashion Store Concept Design Tool

*Paolo Cremonesi, Franca Garzotto, Marco Gribaudo,
Pietro Piazzolla, Mauro Iacono527*

A Software Framework For Intelligent Computer-Automated Product Design

*Robin T. Bye, Ottar L. Osen, Birger Skogeng Pedersen,
Ibrahim A. Hameed, Hans Georg Schaathun.....534*

On Usage Of EEG Brain Control For Rehabilitation Of Stroke Patients

*Tom Verplaetse, Filippo Sanfilippo, Adrian Rutle, Ottar L. Osen,
Robin T. Bye544*

A Game-Based Learning Framework For Controlling Brain-Actuated Wheelchairs

*Rolf-Magnus Hjorungdal, Filippo Sanfilippo, Ottar L. Osen,
Adrian Rutle, Robin T. Bye554*

Intelligent Computer-Automated Crane Design Using An Online Crane Prototyping Tool

*Ibrahim A. Hameed, Robin T. Bye, Ottar L. Osen,
Birger Skogeng Pedersen, Hans Georg Schaathun564*

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems - Special Session -

Big Data As A Service For Monitoring Cyber-Physical Production Systems <i>Alessandro Marini, Devis Bianchini.....</i>	579
Atomic Instruction Translation Towards A Multi-Threaded QEMU <i>Alvise Rigo, Alexander Spyridakis, Daniel Raho.....</i>	587
Towards Secure Non-Deterministic Meta-Scheduling For Clouds <i>Agnieszka Jakobik, Daniel Grzonka, Joanna Kolodziej, Horacio Gonzalez-Velez</i>	596
The Model Of Data Delivery From The Wireless Body Area Network To The Cloud Server With The Use Of Unmanned Aerial Vehicles <i>Ruslan Kirichuk.....</i>	603
Simulation Of Robot-Assisted WSN Localization Using Real-Life Data <i>Michal Marks, Ewa Niewiadomska-Szynkiewicz.....</i>	607
Performance Evaluation Of SOA In Clouds <i>Ashraf M. Abusharekh, Alexander H. Levis</i>	614
Three Layers Network Influence On Cloud Data Center Performances <i>Marco Gribaudo, Mauro Iacono, Daniele Manini.....</i>	621
A Multi-Formalism Framework To Generate Diagnostic Decision Support Systems <i>Giuseppe Cicala, Marco De Luca, Marco Oreggia, Armando Tacchella</i>	628
Characterizing Web Sessions Of E-Customers Interested In Traditional And Innovative Products <i>Grazyna Suchacka, Grzegorz Chodak.....</i>	635
Cloud Implementation Of Agent-Based Simulation Model In Evacuation Scenarios <i>Andrzej Wilczynski, Joanna Kolodziej.....</i>	641

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

Simulation And Selection Of Efficient Decision Rules In Bank's Manual Underwriting Process

Mikhail Konovalov, Rostislav Razumchik.....651

Statistical Classification In Monitoring Systems

Alexander A. Grusho, Nick A. Grusho, Elena E. Timonina658

Two-Sided Truncations Of Inhomogeneous Birth-Death Processes

Yacov Satin, Anna Korotysheva, Ksenia Kiseleva, Galina Shilova, Elena Fokicheva, Alexander Zeifman, Victor Korolev663

Asymptotic Expansions For The Distribution Function Of The Sample Median Constructed From A Sample With Random Size

Vladimir E. Bening, Victor Korolev, Alexander Zeifman669

Uniform In Time Bounds For “No-Wait” Probability In Queues Of $M_t/M_t/S$ Type

Alexander Zeifman, Anna Korotysheva, Yacov Satin, Galina Shilova, Rostislav Razumchik, Victor Korolev, Sergey Shorgin.....676

Hybrid Simulation Of Active Traffic Management

Anna V. Korolkova, Tatyana R. Velieva, Pavel Abaev, Leonid A. Sevastianov, Dmitry S. Kulyabov.....685

SIR Analysis In Square-Shaped Indoor Premises

Andrey Samuylov, D. Moltchanov, Yu. Gaidamaka, V. Begishev, R. Kovalchukov, Pavel Abaev, Sergey Shorgin692

Stochastization Of One-Step Processes In The Occupations Number Representation

Anna V. Korolkova, Ekaterina G. Eferina, Eugeny B. Laneev, Irina A. Gudkova, Leonid A. Sevastianov, Dmitry S. Kulyabov698

On Analytical Modeling Of IMS Conferencing Server

Pavel Abaev, Vitaly Beschastny, Alexey Tsarev.....705

New Scheduling Policy For Estimation Of Stationary Performance Characteristics In Single Server Queues With Inaccurate Job Size Information

Lusine Meykhanadzhyan, Rostislav Razumchik710

Author Index	717
---------------------------	-----

ECMS 2016

SCIENTIFIC PROGRAM

Plenary Talks

Discrete Event Optimization: Theory, Applications And Future Challenges

Andrea Matta
Shanghai Jiao Tong University
School of Mechanical Engineering, Dept. of Industrial Engineering and Management
800 Dong Chuan Road, Shanghai
200240, P.R. China
E-mail: matta@sjtu.edu.cn

ABSTRACT

Optimization of discrete event systems is often time consuming and also requires specific approaches due to the fact that general methodologies cannot be successfully applied to any kind of system. Conventional approaches use simulation as a black-box oracle to estimate performance at design points generated by a separate optimization algorithm. This decoupled approach fails to exploit an important advantage: simulation codes are white-boxes, at least to their creators. In fact, the full integration of the simulation model and the optimization algorithm is possible in many situations.

The methodology Discrete Event Optimization (DEO) is presented. DEO allows the development of integrated simulation-optimization models for queueing systems by means of the ERGLite formalism, a subclass of ERGs (Entity Relationships Graphs). Furthermore, DEO provides a formal way to map ERGLs into mathematical formulations for optimization of queueing systems. In case the obtained model is a MILP (Mixed Integer Linear programming), DEO also provides a formal way to approximate the obtained models based. The analytical properties of the obtained models are analyzed in the frameworks of Sample Path Optimization and Mathematical Programming. Several examples will be presented to show the applicability of DEO and to point out its strengths and drawbacks. Research challenges will also be identified.

Data Analytics, Model Generation And Optimization Algorithms A Perfect Match?

Thomas Hußlein
OptWare GmbH
Prüfeninger Straße 20
93049 Regensburg, Germany
www.optware.de - info@optware.de

ABSTRACT

To provide a timely and cost-effective reaction to the ever changing planning tasks within production and logistics, automated planning and optimization methods gain more and more acceptance with industrial applications. Every OR-based solution for production- and logistics planning requires a mathematical model of the relations of the different parameters and variables. Presently the creation of the model is performed by human experts. Due to the complexity and high frequency of changes within the logistics and productions processes, a detailed modeling for these processes by humans often is not possible or is too costly. In the approach presented here a robust model with good accuracy and reduced complexity is created automatically by data analysis.

The result is the prediction of the systematic behavior of logistics processes that allows to keep the model up to date at almost no additional cost. Subsequently the obtained model is used as an input for automated optimization algorithms. The presented approach combines methods from Data Analysis, Artificial Intelligence and Mathematical Optimization. An application for car manufacturing processes is provided. The prospects for the generalized application in many environments are outlined.

Agent-Based Simulation

A LEARNING AGENT FOR A MULTI-AGENT SYSTEM FOR PROJECT SCHEDULING IN CONSTRUCTION

Florian Wenzler
Willibald A. Günthner
Lehrstuhl für Fördertechnik Materialfluss Logistik
Technische Universität München
Boltzmannstraße 15, 85748 Garching, Germany
Email: wenzler@fml.mw.tum.de, guenther@fml.mw.tum.de

KEYWORDS

MRCPSP; Project Scheduling, Multi-agent system, Discrete event simulation

ABSTRACT

The quality of the project plan created is essential for realizing a construction project. This is a big challenge for planners, because there are many constraints to be considered. The problem to be solved is known as the multi-mode resource-constrained project scheduling problem (MRCPSP). This paper presents a multi-agent approach in which resources and processes are represented as collaborative agents. Autonomous process and resource agents register themselves on a central blackboard where resource allocation to activities is negotiated. As expansion to prior works, a learning agent is integrated to improve the solutions created. A discrete-event simulation implements the model and it is evaluated with standardized project plans from the field of operations research.

INTRODUCTION

Insufficient construction-project planning often leads to overall progress delays and cost overruns. Most projects are nonetheless scheduled manually without using optimization tools. Hence, quality depends on the planner's experience and the available time. That's why project plans are often generated without much detail or consideration of constraints, which are primarily predecessor/successor dependencies, and limited resources and space.

The influence of unpredictable circumstances is also important for project management in construction. The former lead to delays and necessitate rescheduling, but the effect of delayed processes cannot be investigated in advance in detail without a computer-based tool.

Most project scheduling software uses methods such as Program Evaluation and Review Technique (PERT) or Critical Path Method (CPM) (Maroto and Tormos 1994). None of those methods considers resource constraints. A method for project scheduling in construction dealing with these topics therefore has to be developed, which

- considers all types of constraints in construction,
- is adaptable to specific situations, and
- enables easy rescheduling after unpredictable incidences.

PROBLEM STATEMENT

The general problem to be solved is known as the resource-constrained project scheduling problem (RCPS) or the multi-mode resource-constrained project scheduling problem (MRCPSP), because every activity can be executed in different ways (modes). These modes' process time and required resources differ. The problem can be described generally as follows:

- J: number of activities/jobs
- j: activity ID with $j = \{0, \dots, J+1\}$
- M: number of modes for each activity j
- d_{jm} : duration of activity j executed in mode $m \in M$
- S_j : successors of activity j
- P_j : predecessors of activity j
- R: number of different types of renewable resources
- N: number of different types of nonrenewable resources
- r_{jmk} : renewable resources of type $k \in R$ required by activity j in mode m
- n_{jml} : nonrenewable resources of type $l \in M$ required by activity j in mode m

Jobs having IDs 0 and J+1 are dummy activities with neither a processing time nor resource requirements ($d_{0\ m|J+1\ m}=0$, $r_{0\ mk|J+1\ mk}=0$, and $n_{0\ ml|J+1\ ml}=0$). They serve as the project's start and end.

Minimizing a project's makespan while taking care of the given constraints is the goal. The following variances can be used to achieve this: first is the mode in which an activity executes; second is each process's starting time. That the starting time's predecessor/successor dependencies aren't violated has to be guaranteed.

The number of resources used in the project plan created is never allowed to exceed the given number of renewable and nonrenewable resources. The chosen solution is invalid if it does so.

Some simplifications have to be made for upcoming parts of the paper:

- The execution of started activities cannot be interrupted.
- An activity's chosen mode cannot be subsequently changed.
- An activity's resources remain assigned until the job is finished.
- The number of available resources cannot be changed during the project time.

However for the intended use in construction, these restrictions have no big influence or can be considered by adjusting the input data (e.g., splitting an activity up into two or more parts with individual features).

Different schedules are needed during development and for tests. Kolisch and Sprecher created standardized examples for this purpose with their project generator, ProGen (Kolisch and Sprecher 1997). The plans fulfill all constraints mentioned and are built up systematically for selected parameters as Table 1 shows. A particular parameter is changed for every type of plan while the rest remain fixed. This allows selective investigation of each parameter's influence on the result.

Table 1: Structure of the Project Plans Used

Name	Parameter			
	J	M	R	N
j10	10	3	2	2
j16	16	3	2	2
j30	30	3	2	2
m1	16	1	2	2
m5	16	5	2	2
n0	10–20	3	2	0
n3	16	3	2	3
r1	16	3	1	2
r5	16	3	5	2

Knowledge of the minimal project duration is an essential advantage of the instances that Kolisch and Sprecher created. The quality of the method used to solve the MRCPSP can thus be compared and evaluated.

DIFFERENT APPROACHES TO SOLVING THE RCPSP AND MRCPSP

Blazewicz et al. proved that this problem is np-hard (Blazewicz et al. 1983). An optimal solution is hence nearly impossible to find within a reasonable amount of time. For smaller projects, approaches such as branch-and-bound (Johnson 1967) or lower bounds (Heilmann and Schwindt 1997) can be used to find the optimum. However, the solution space grows very fast with larger projects and these approaches become inefficient. That's why various heuristics and meta-heuristics were developed and adapted for the (M)RCPSP. Among these

are simulated annealing (König and Beißert 2009, Józefowska et al. 2001), genetic algorithms (van Peteghem and Vanhoucke 2010, Senouci and Al-Derham 2008, Toklu 2002), ant colony algorithms (Li and Zhang 2013, Christodoulou 2005), or particle swarm optimization (Jarboui et al. 2008, Lu et al. 2008, Zhang et al. 2006). Despite their different basic ideas, they all create new combinations according to different rules, but also randomly, to try to find a better solution.

Since creating every possible solution within an acceptable time isn't possible, finding the optimal solution it is not guaranteed.

MULTI-AGENT SYSTEM FOR THE MRCPSP

Using a Multi-Agent System (MAS) is a different approach to solving this problem. The main benefit is being able to split the whole problem into smaller, easier parts. Furthermore, the latter are more robust and flexible than those in traditional methods (Davidsson et al. 1994). The agents themselves are also easy to understand and create due to the small number of capabilities.

A few MAS implementations exist for the resource-constrained scheduling problem. Horenburg presented a MAS for the RCPSP with agents for each activity as well as for each resource. Resource allocation to jobs is controlled by priority rules (Horenburg 2014). Knotts et al. introduced another agent-based framework for solving the RCPSP in minimal project duration (Knotts et al. 2000). Resources aren't modeled as agents in this case.

Wauters et al. implemented a new aspect with the system's ability to learn (Wauters et al. 2011). Selecting the next activity is realized in two steps for solving the MRCPSP. The most important job is first identified, then one of its modes is chosen. This means consequently that not all modes of the activities have the same chance to get executed. If the mode of an activity with the highest priority cannot be executed, it is not possible to select a mode from an alternative activity, although it might be a better choice than the next mode from the previously chosen activity. With this issue deals the following presented MAS by including all possible modes in the process of resource allocation. Hence, only one step is needed for choosing the next mode of an activity and there might be potential for improvements of a Multi-Agent System.

Framework of the MAS

This section will present the structure of a multi-agent system for the MRCPSP. Figure 1 shows the different types of agents and communication. Different types of agents represent activities as well as renewable and nonrenewable resources. The central element is the blackboard. Resource allocation to current activities is negotiated there. This architecture simplifies communication (compared to the complexity required

when all agents have to communicate with each other) and promotes efficient, transparent resource allocation.

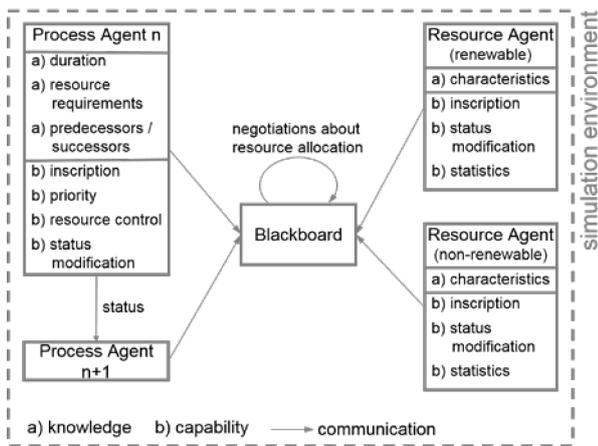


Figure 1: Multi-Agent System

All of the jobs' executable modes register themselves on the blackboard with resource requests as soon as their predecessors finish. All available resources are simultaneously inscribed there too. Every time a job finishes, this agent informs all the former's successors about the completion.

Information about duration, resource requirements, and previous/subsequent activities enables every process agent to act in the described way. Beside these characteristics, every agent can act on and communicate with its environment. As already mentioned, the most important capability is being able to register on the blackboard with the calculated priority value. Methods also exist for adjusting status and recording data for statistics.

Priority Rules for Resource Allocation

Insufficient resources typically exist for all of activities registered on the blackboard. To identify the most crucial current jobs, each process agent transmits a priority value. The activities' negotiation order is calculated based on this value. Whether enough resources are available at the moment is successively checked for each. If not, an activity is postponed until the next negotiation round. That the limits of simultaneously active renewable resources are never exceeded can be guaranteed this way.

The situation is different with nonrenewable resources. Subsequent activities cannot be started once the limit is reached, and the search for a solution stops prematurely. Due to the way the project plan is created, a valid combination of modes is not guaranteed. That an early negotiation can cause too many nonrenewable resources to be used is unavoidable with local decisions. The project is nevertheless planned completely for getting an (invalid) starting combination, which can be improved later.

Different priority rules for the MAS introduced to solve the MRCPSp were presented in a previous paper (Wenzler and Günthner 2015). They feature different activity attributes to compute the priority value such as duration, resource requirements, or number of successors.

The LPF_AVG (Longest Path Following) rule is chosen in the sequel. This was shown to provide - together with others - the best results and is defined as follows: Every activity determines the duration of its successor processes. The activity with the biggest value receives the highest priority. Since priority calculation occurs before or during project planning itself, the longest path has to be identified without resources. Appendix “_AVG” defines how to handle the different modes of every activity in the path. Every activity can be executed in only one mode, but which one will be chosen is unknown in advance. So the average of all modes is assumed for an activity's duration in this priority rule.

Introducing a Learning Agent

As mentioned, the first simulation run may be unable to find the optimal solution. That's why a new agent type, the learning agent (LA), was incorporated into the existing framework. Figure 2 shows its communication with other agents.

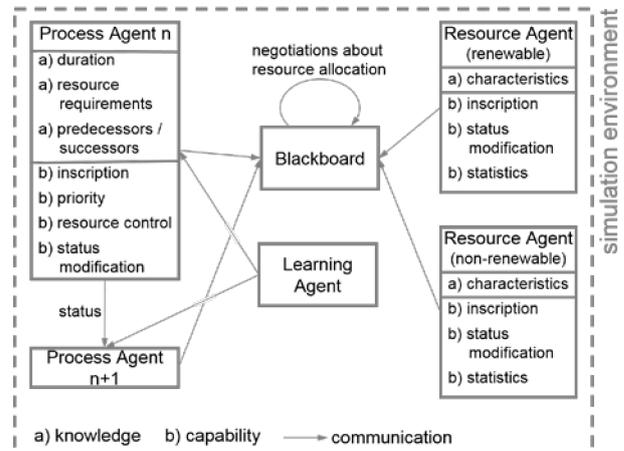


Figure 2: Multi-Agent System with Learning Agent

The learning agent can analyze the plan created so far and influence the process agents' mode choices. The LA subsequently restarts the planning procedure and compares the result with previous solutions.

THE LEARNING AGENT'S FUNCTIONALITY

The learning agent (LA) has two main tasks, which are executed in the order listed:

- Create a feasible solution that doesn't exceed resource limits.

- Improve a feasible solution as far as possible.

The LA is active until a stop criterion is satisfied. This can happen in several different ways:

- The optimum is found. Therefore, the best solution has to be known. Projects are usually so large that the optimum cannot be determined. For plans from the PSPLIB, which are used in this paper, the minimum makespan for each schedule is known and the LA can use this knowledge.
- One type of optimization rule is used successively more often than a defined limit.
- No improvements are made for too many consecutive times.

With the last two rules, the calculation will terminate whenever the LA cannot improve the solution with the defined settings.

Creating Feasible Solutions

A valid solution - even one with a longer makespan - is better than exceeding the constraints. Hence, the LA's first task is to generate a feasible plan.

The heuristic of the learning agent to get an acceptable solution operates as follows: An (invalid) solution is needed first. Then each activity's modes are cycled through to check for possible nonrenewable-resource improvements. The LA obeys the following rules to do this:

- The requirement for at least one type of nonrenewable resource is less stringent in the new mode than in the current one.
- The amounts of other types of resources used must not grow - unless enough reserve exists.
- The limits imposed on other types of resource may not be exceeded.

The result of this procedure depends on the initial solution. A new combination of modes will be chosen if a feasible solution was not generated. The mode with the highest savings is selected to avoid the previous bottleneck for the crucial type of nonrenewable resources.

As soon as a feasible solution is found, the LA transmits the defined modes to the process agents and the process of creating a schedule is started again.

Solution Improvement

Any feasible solution generated is unlikely to be the optimal solution. The LA is hence tasked with improving it via selective adjustments. Changing the mode, the earliest starting time, and the priority rule are possible adjustments.

One important solution analysis tool is identification of the critical path and the floats. If the value of an

activity's float is greater than 0, this activity can be delayed up to this value without having any influence on the remaining schedule. Every activity with zero float is part of the critical path. With this information, searching the activities the adjustment of which probably effects the schedule most is possible. The following rules are implemented in the current state of research:

- Change some activities to a mode with shorter duration. The number of selected activities can vary. For every activity the ratio of duration to resource requirement is calculated and for those activities, whose value is above the average, a new mode is chosen.
- Shift some of the activities to a later start time in case they have enough float. The freed resources may allow other activities to start earlier.
- Execute some activities in a mode with less demanding resource requirements so other jobs can use more resources or start earlier. Chosen are those activities which save more resources than the average by changing the mode.

At this point, no rule uses random for changing the parameters. For that reason, every decision is understandable.

RESULTS

The MAS presented was implemented in a discrete-event simulation (DES). Monte-Carlo simulations were conducted to verify and validate the model as well as to provide reference values. Priority values are therefore generated randomly. Several different priority rules were evaluated in the next step (Wenzler and Günthner 2015). The "LPF_AVG" rule produced the best results so this rule was chosen in this paper (labeled "without LA" in tables or figures).

Comparison with the MAS without LA

This section will present the effect of activating the learning agent on the simulation results. The first goal for which the learning agent is implemented is to reduce the number of invalid schedules. Table 2 lists the number of projects for which no feasible solution was found.

Table 2: Number of Infeasible Projects

Type	Total number of projects	Infeasible solutions	
		Without LA	With LA
j10	536	315	0
j16	550	308	0
j30	552	300	0
m1	640	0	0
m5	558	309	4
n0	470	0	0
n3	600	372	15
r1	553	306	0
r5	546	286	0

Without the active LA for every type of plan, the MAS left a number of projects unsolved. The ratio is up to 62% except for m1 and n0, where all plans are solvable because of their structure.

Table 2 shows that a valid schedule was created using the LA for almost every project. The only exceptions are the most complex plans, m5 and n3, with 4 and 15 unsolved plans respectively. All possible combinations of modes for each activity were searched by enumeration to further investigation of why the MAS with LA still cannot solve some of the projects (Table 3). The “Feasible combinations” column represents the number of different combinations that can be created without exceeding nonrenewable-resources limits.

Table 3: Number of Feasible Combinations of the Unsolved Schedules

Type	Number	Feasible combinations	Possible combinations
n3	1 6	1	43 046 721
	3 3	189	43 046 721
	3 6	27	43 046 721
	6 3	4	43 046 721
	6 4	6	43 046 721
	6 6	1	43 046 721
	6 7	4	43 046 721
	6 8	4	43 046 721
	7 7	4	43 046 721
	8 2	16	43 046 721
	8 4	4	43 046 721
	36 7	18	43 046 721
	36 8	124	43 046 721
	36 9	1881	43 046 721
	36 10	2	43 046 721
m5	1 1	4	1 440 000 000
	1 2	2	35 156 250 000
	5 4	256	152 587 890 625
	36 9	1104	152 587 890 625

The results show that are only a few possible ways exist to get a feasible schedule. Sometimes the heuristic has to find the single way out of more than 43×10^6 possibilities, as in case of the n3 plans, or one of two solutions from 35×10^9 combinations theoretically possible for project-type m5.

The heuristics for solving these projects correctly have to be improved in future work. Integrating enumeration is not an option because of excessive computing time especially for large projects.

Table 4 shows the results of the LA’s second task: solution improvement. The number of optimal solutions increased only slightly with the defined stop criteria for plan types j16, n3, and r1. However, a lot of the remaining projects finished within a shorter time.

Table 4: Project-Makespan Improvement

Type	Optimal solutions		Better solutions with LA
	without LA	with LA	
j10	112	112	351
j16	112	113	360
j30	116	116	366
m1	400	400	0
m5	96	96	305
n0	231	231	88
n3	107	110	393
r1	136	137	337
r5	136	136	353

The figures below show detailed results for some project types. The number of tested schedules having a certain deviation from the known optimum can be seen there. The bar with “0” deviation represents the optimal solutions, while the declared value of time units is also needed for completion of the other plans. The last bar shows the number of infeasible solutions if any exist.

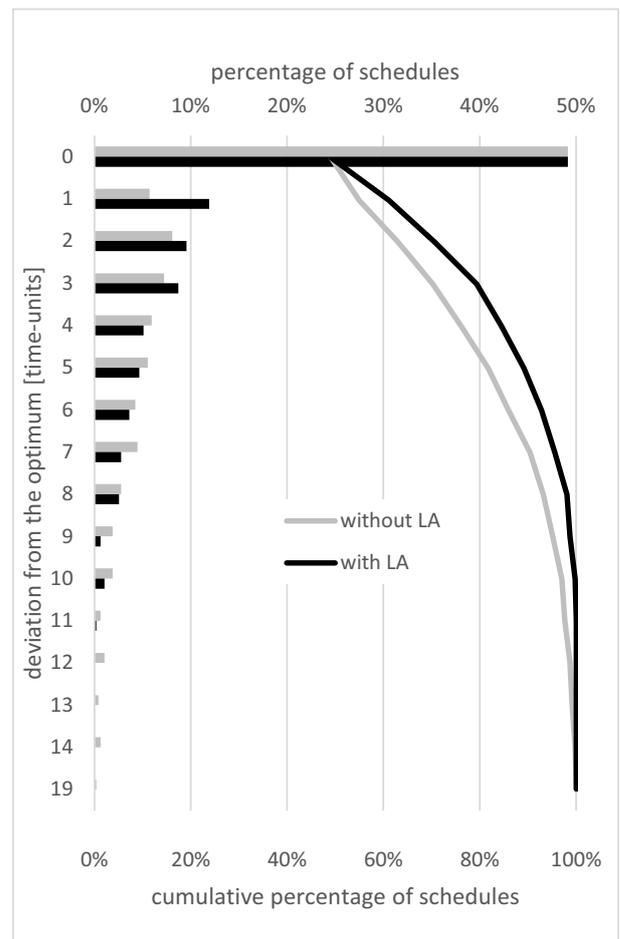


Figure 3: Detailed Results for Project Type n0

Nearly 50% of the plans specified n0 can be solved optimally with the MAS (Figure 3). This can be achieved even without the LA; however, the other

results improve with the active LA. The largest deviation is reducible from 19 to 11 time units. The number of nearly perfect solutions with deviations of 1 to 3 also rose significantly.

The n3 projects' greater complexity is visible in Figure 4. More than 60% irregular schedules exist without the LA. In contrast, only 15 unsolved projects remain with the LA. The number of optimal solutions or of those with small deviations from the optimum also increased. The main point for further improvements is the large number of schedules for which the heuristics found a feasible but not optimal solution.

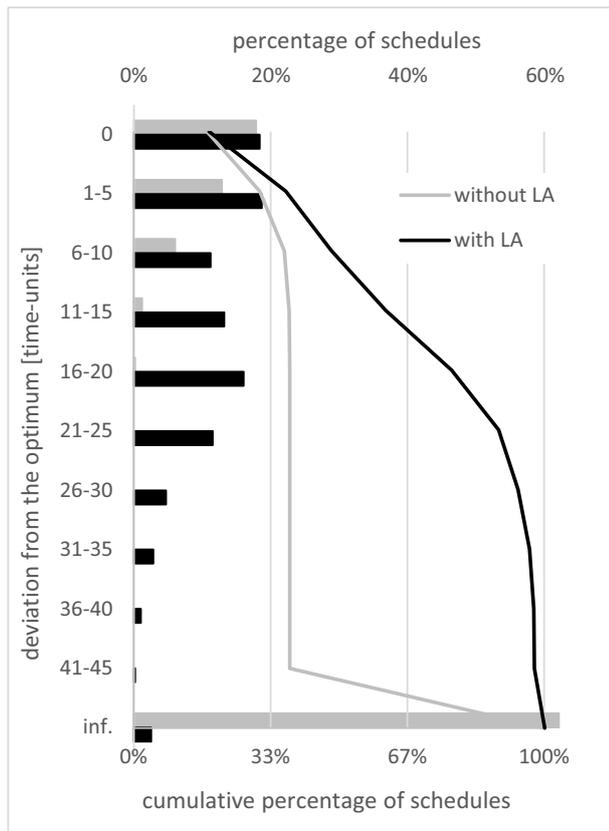


Figure 4: Detailed Results for Project Type n3

Comparison with other Approaches

In Table 5 the results of the MAS with the priority rule “LPF_AVG” for the datasets with 10 to 20 activities (j10-j20) are shown. For those, the comparison with

Table 5: Comparison with other Approaches for the MRCPSPP – Average Deviation from Optimum (%)

	j10	j12	j14	j16	j18	j20
LPF AVG	41.95	41.44	43.05	44.45	45.16	45.48
Li and Zhang (2013)	0.09	0.13	0.40	0.57	1.02	1.10
Wauters et al. (2011)	0.05	0.08	0.23	0.30	0.53	0.70
Van Peteghem and Vanhoucke (2010)	0.01	0.09	0.22	0.32	0.42	0.57
Jarboui et al. (2008)	0.03	0.09	0.36	0.44	0.89	1.10
Józefowska et al. (2001)	1.16	1.73	2.60	4.07	5.52	6.74

other approaches is possible. The table shows the average deviation from the known optimal solution. The actual performance of the MAS is not as good as those of the alternative methods. This can be explained by the following issues. Firstly, the LA creates feasible solutions without giving the project duration top priority. These solutions have in general a large makespan (up to 200% of the optimal duration) and so even a few solutions with a long duration have a strong influence on the average deviation. To solve this problem, the LA has to improve the initial solution by changing some parameters. In the current state, only the mentioned rudimental rules are implemented and after about 10 iterations the solutions aren't changing anymore. Therefore the heuristics have to be improved to create a larger solution space.

The positive aspect of the actual results is, that the average deviation is nearly constant, although the size of the datasets increases.

CONCLUSION

This paper presents a multi-agent approach to solving the MRCPSPP. An individual collaborative agent, a new type of which (learning agent) was introduced, represents every activity and resource. It analyses a previously generated solution and influences the process agents' decisions concerning the chosen mode using the dependent resource requirements or the starting time.

The MAS is implemented in a discrete event simulation environment and tested with standardized projects from the field of operations research. Hence, these projects' optimal solution is known, and the quality of the project plans created could be evaluated.

With the learning agent (LA), the high quota of irregular project plans can be reduced significantly and the number of (nearly) optimal solutions increased compared to the MAS without learning agent.

The presented MAS is a preliminary result. The learning agent has to be improved further to create better solutions with as little rescheduling as possible by the end of the project.

Some specific additions have to be made for use in construction. First, a new area agent allows the limited space on a construction to be taken into account site. That leads to new constraints, which have to be considered.

A type of resource agent for shared resources is missing. Several processes use some machines simultaneously (e.g., cranes) necessitating another agent. Include the emerging interactions between participating activities in the MAS is possible this way.

Finally, real project data will be used to demonstrate applicability.

ACKNOWLEDGMENTS

The research presented in this paper is part of the project "Entwicklung einer agentenbasierten Methodik zur Terminplanoptimierung im Bauwesen unter Berücksichtigung ressourcenabhängiger Prozesslängen" (GU 427/22-1), which is funded by the German Research Foundation (DFG).

REFERENCES

- Blazewicz, J.; Lenstra, J. and Rinnooy Kann, A. H. G. 1983. "Scheduling subject to resource constraints: Classification and complexity." *Discrete Applied Mathematics* 5, No. 1 (Jan), 11–24.
- Christodoulou, S. 2005. "Ant Colony Optimization on Construction Scheduling." In *Proceedings of the International Conference on Computing in civil engineering 2005* (Cancun, Mexico, July 12–15), American Society of Civil Engineering.
- Davidsson, P.; Astor, E. and Ekdahl, B. 1994. "A framework for autonomous agents based on the concept of anticipatory systems." In *Proceedings of Cybernetics and Systems '94*. World Scientific, Singapore, 1427–1434.
- Heilmann, R. and Schwindt, C. 1997. "Lower Bounds for RCPSP/max." Report WIOR-511. Universität Karlsruhe.
- Horenburg, T. 2014. *Simulationsgestützte Ablaufplanung unter Berücksichtigung aktueller Baufortschrittsinformationen*. Dissertation Technische Universität München, Lehrstuhl für Fördertechnik Materialfluss Logistik.
- Jarboui, B.; Damak, N.; Siarry, P. and Rebai, A. 2008. "A combinatorial particle swarm optimization for solving multi-mode resource-constrained project scheduling problems." *Applied Mathematics and Computation* 195, No. 1 (Jan), 299–308.
- Johnson, T.J.R. 1967. *An algorithm for the resource-constrained project scheduling problem*. Ph.D. Dissertation. MIT.
- Józefowska, J.; Mika, M.; Rózycki, R.; Waligóra, G. and Weglarz, J. 2001. "Simulated annealing for the multi-mode resource-constrained project scheduling." *Annals of Operations Research* 102, No. 1 (Feb), 137–155.
- Knotts, G.; Dror, M. and Hartmann, B. C. 2000. "Agent Based project scheduling." *IIE Transactions* 32, No. 5 (May), 387–401.
- Kolisch, R. and A. Sprecher. 1997. "PSPLIB – A project scheduling problem library." *European Journal of Operational Research* 96, No. 1 (Jan), 205–216.
- König, M. and Beißert, U. 2009. "Construction Scheduling Optimization by Simulated Annealing." In *Proceedings of the 26th Annual International Symposium on Automation and Robotics in Construction* (Austin, TX, June 24–27). International Association for Automation and Robotics in Construction, 183–190.
- Li, H. and Zhang, H. 2013. "Ant colony optimization-based multi-mode scheduling under renewable and nonrenewable resource constraints." In *Automation in Construction* 35 (Aug), 431–438.
- Lu, M.; Lam, H.-C. and Dai, F. 2008. "Resource-constrained critical path analysis based on discrete event simulation and particle swarm optimization." In *Automation in Construction* 17, No. 6 (Nov), 670–681.
- Maroto, C. and Tormos, P. 1994. "Project Management: an Evaluation of Software Quality". In *International Transactions in Operational Research*, 1, No. 2 (Apr), 209–221.
- Senouci, A. and Al-Derham, H. R. 2008. "Genetic algorithm-based multi-objective model for scheduling of linear construction projects." In *Advances in Engineering Software* 39, No. 12 (Dec), 1023–1028.
- Toklu, Y.C. 2002. "Application of genetic algorithms to construction scheduling with or without resource constraints." In *Canadian Journal of Civil Engineering* 29, No. 3 (Jun), 421–429.
- Van Peteghem, V. and Vanhoucke, M. 2010. "A genetic algorithm for the preemptive and non-preemptive multi-mode resource-constrained project scheduling problem." *European Journal of Operational Research* 201, No. 2 (Mar), 409–418.
- Wauters, T.; Verbeeck, K.; Vanden Berghe, G. and De Causemaecker, P. 2011. "Learning agents for a multi-mode project scheduling problem." *Journal of the Operational Research Society* 62, 2 (Feb), 281–290.
- Wenzler, F. and W. A. Günthner. 2015. "Ressourcenbeschränkte Terminplanung mit einem System kollaborativer Agenten." In *Simulation in Production and Logistics 2015*, M. Rabe and U. Clausen (Eds.). Fraunhofer, Stuttgart, 721–730.
- Zhang, H.; Tam, C. M. and Li, H. 2006. "Multimode Project Scheduling Based on Particle Swarm Optimization." In: *Computer-Aided Civil and Infrastructure Engineering* 21, No. 2 (Feb), 93–103.

FLORIAN WENZLER studied Mechanical Engineering with concentration on automotive and power train technologies at the Technische Universität München (TUM). He obtained his degree in 2013 and has been working since then as a research assistant at the Institute for Materials Handling, Material Flow, Logistics (fml). His work focuses on operations research and the discrete event simulation of logistic systems. His email address is: wenzler@fml.mw.tum.de.

WILLIBALD A. GÜNTNER is professor and head of the Institute for Materials Handling, Material flow, Logistics (fml) at the Technische Universität München (TUM) He is cofounder of the Wissenschaftliche Gesellschaft für Technische Logistik e. V. and advisory board member of several associations, federations, and companies.

AGENT-BASED MODEL CONTINUITY OF STOCHASTIC TIME PETRI NETS

Franco Cicirelli¹, Libero Nigro², Paolo F. Sciammarella²

¹CNR - National Research Council of Italy
Institute for High Performance Computing and Networking (ICAR) - 87036 Rende(CS) - Italy
²Software Engineering Laboratory
University of Calabria, DIMES - 87036 Rende (CS) – Italy
Email: f.cicirelli@dimes.unical.it, l.nigro@unical.it, p.sciammarella@dimes.unical.it

KEYWORDS

Multi-agent systems, model continuity, simulation, real-time, stochastic time Petri nets, Java, JADE.

ABSTRACT

Stochastic Time Petri Nets (sTPN) are a useful formalism for modelling and quantitative analysis of concurrent systems with timing constraints. This paper describes an implemented tool supporting sTPN, which was achieved on top of a control-centric agent-based framework which fosters model continuity. Model continuity means the same model can be used for property checking through simulation and for real-time execution. The paper demonstrates the effectiveness of the approach through a modelling example.

INTRODUCTION

Stochastic systems can be studied by either numerical or statistical solution techniques (Younes et al., 2006). Numerical methods enumerate the stochastic states of a model and can evaluate a probability measure over a path of state transitions by solving equations based on the state associated probability distribution functions. Numerical methods tend to be more accurate than statistical methods which are based on sampling and simulation. However, numerical methods can suffer of state explosion problems and can impose restrictions on the classes of modelled systems, e.g., based on timers which satisfy the Markov property or which admit regeneration points in more general systems. Stochastic Time Petri Nets (sTPN) (Paolieri et al, 2016) have been proposed for modelling and analysis of concurrent systems with timing constraints. They are supported by numerical techniques in the context of the ORIS tool (Bucci et al., 2010). An approach to statistical model checking of sTPN based on UPPAAL is described in (Cicirelli et al., 2015). This paper proposes an original agent-based tool supporting sTPN. Novel in the tool is a support to *model continuity* (Cicirelli&Nigro, 2016a-b) that is the possibility of using a same model for temporal analysis by simulation and for real-time execution. The paper first describes the definitions of sTPN. Then a summary of the underlying control-centric agent-based architecture is furnished. After that an overview of the tool implementation is provided. The developed approach is then demonstrated by a case study concerning a probabilistic formulation of

the Fisher’s mutual exclusion algorithm (Lynch&Shavit, 1992)(Paolieri et al., 2016). Finally, conclusions are presented with an indication of on-going and future work.

STOCHASTIC TIME PETRI NETS

Syntax

An sTPN is a tuple

$$(P, T, B, F, M_0, I_{nh}, E, Uw, Ud, EFT^s, LFT^s, PDF, W)$$

where:

- P and T are disjoint finite nonempty set of places and transitions; $T = T_i \cup T_t$ where T_i are immediate transitions, and T_t are timed transitions;
- B is the backward incidence function, $B: P \times T \rightarrow \mathbb{N}$, where \mathbb{N} denotes the set of natural numbers;
- F is the forward incidence function, $F: P \times T \rightarrow \mathbb{N}$;
- M_0 is the initial marking function, $M_0: P \rightarrow \mathbb{N}$, which associates with each place a number of tokens;
- I_{nh} is the set of inhibitor arcs, $I_{nh} \subset P \times T$ where $(p, t) \in I_{nh} \Rightarrow B(p, t) = 0$;
- $E: T \rightarrow \{true, false\}$ is a boolean function which extends the enabling condition of a transition. If omitted, it defaults to *true*;
- Uw and Ud are two update functions which extend respectively the withdraw/deposit phase of a transition. If omitted, they default to *void*;
- $EFT^s: T_t \rightarrow R^+$ is a function which associates each timed transition with a (finite) earliest static firing time. R^+ denotes the set of non-negative real numbers;
- $LFT^s: T_t \rightarrow R^+ \cup \{\infty\}$ is a function which associates each timed transition with a (possibly infinite) latest static firing time. It must be $LFT^s \geq EFT^s$. An immediate transition logically has $EFT^s = LFT^s = 0$;
- PDF is a function which associates each timed transition with a probability distribution function constrained in the interval $[EFT^s, LFT^s]$;
- W is a function, $W: T_i \rightarrow R^+$, which associates each immediate transition with a weight.

Semantics

A transition t is *enabled* if each of its input places contains sufficient tokens and $E(t)$ evaluates to *true*, i.e., iff

$$\forall p \in P, (p, t) \in I_{nh} \Rightarrow M(p) = 0 \wedge B(p, t) > 0 \Rightarrow M(p) \geq B(p, t) \wedge E(t)$$

An enabled immediate transition t_i is *fireable*. Fireability of immediate transitions always has priority over that of timed transitions. Among the set of fireable immediate transitions, each t_i can *fire* with probability

$$Prob(t_i) = \frac{W(t_i)}{\sum_{t_j \in T_i \text{ and } t_j \text{ is enabled}} W(t_j)}$$

The *time-to-fire* $\tau(t_t)$ of a timed transition t_t is stochastically defined, at its enabling instant, by sampling its associated *PDF*(t_t) with the constraint:

$$EFT^s(t_t) \leq \tau(t_t) \leq LFT^s(t_t)$$

A timed transition is fireable at its absolute time-to-fire, i.e., *enabling time*(t_t) + $\tau(t_t)$, provided it is less than or equal to the absolute time-to-fire of all the other simultaneously enabled timed transitions. Timed transitions with the same absolute time-to-fire will fire non deterministically.

Let $m: P \rightarrow N$ be the net marking, which specifies the number of tokens of each place of the sTPN model at a certain instant of time. When the transition t fires, the marking m is replaced by a new marking m' which is derived from m by the withdrawal of tokens from the input places and the deposit of tokens in the output places. More precisely, the firing process consists of the two (atomic) phases:

$$m_{int}(p) = m(p) - B(p, t) - U_w(t) \text{ (withdraw phase)}$$

$$m'(p) = m_{int}(p) + F(p, t) + U_d(t) \text{ (deposit phase)}$$

Transitions which are enabled in m , in the intermediate marking m_{int} and in the final marking m' are said *persistent* to the t firing. Transitions which are enabled in m' but not in m_{int} are said *newly enabled*. Newly enabled timed transitions have their time-to-fire which is resampled.

A transition which is multiple enabled at a time instant is assumed to fire its enablings one at a time (*single server* semantics). Therefore, following its own firing, would t be still enabled, it is regarded as newly enabled.

As a final remark, it should be noted that the functions E , U_w and U_d are model-specific and can be exploited, e.g., for managing a high-level concept like a variable (see Fig. 7) or to avoid cluttering in complex topologies.

CONTROL SENSITIVE AGENT FRAMEWORK

The following highlights the control-based framework (Cicirelli&Nigro, 2016a-b) for building multi-agent systems which is at the basis of the sTPN tool described later in this paper.

The framework is founded on the notions of *actors* (agents) and *actions* (see Fig. 1)

Actors

Actors are modelled as finite state machines which communicate to one another by asynchronous message

passing. Actors are thread-less. They are at rest until a message arrives to be processed. The behavior of an actor (i.e., its state machine) is modelled in its associated *handler()* method. An incoming message causes local variables of the actor to be updated, possibly changes the current state of the state machine, can send new messages to known actors and can submit one or more actions.

A subsystem of actors (Logical Process or LP) is allocated for the execution on a computing node. All the actors of a same subsystem are regulated by a local *control machine* which transparently buffers exchanged messages into one or more message queues and ultimately delivery messages, one at a time, to recipient actors, according to a proper *control strategy*, e.g., based on a time notion (simulated or real-time). Message processing in a actor subsystem represents the *unit of scheduling*.

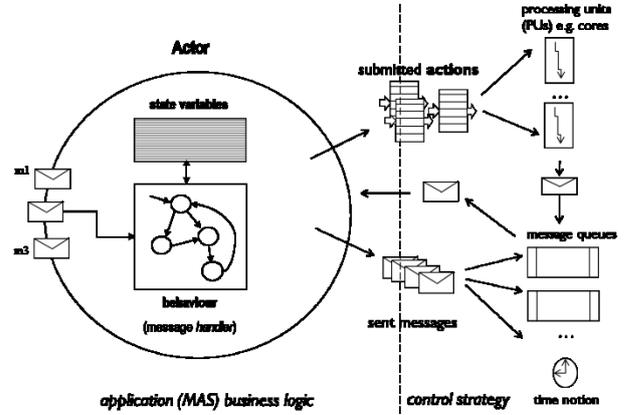


Figure 1 - Actor organization and orthogonal control aspects

In general, multiple actor subsystems can be federated to constitute a distributed system (see Fig. 2), using the services of a suitable transport layer and communication protocol. A *Time Server* can be in charge of maintaining a global time notion across the federated system.

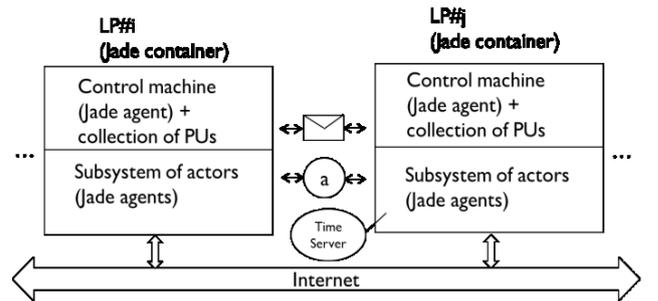


Figure 2 - Federated actor system based on JADE

Fig. 2 portrays a snapshot of a distributed actor system achieved on top of the open source JADE project (Bellifemine et al., 2007)(Cicirelli&Nigro, 2016a-b) which provides basic services for agent lifecycle, naming and message exchanges based on FIPA (Foundation for Intelligent Physical Agents). In addition it favours interoperability with legacy software FIPA compliant. Both actors and messages can be dynamically transferred from an LP (JADE container) to another.

A fundamental design issue of the actor-framework is related to the control machines which act as plug-ins tailored to the application needs.

Actions

Messages promote sociality among actors and capture the occurrence of events. They are handled sequentially in an interleaved way by the local control structure.

Besides messages, the actor framework relies also on *actions*, that are activities which consume time and require *processing units* (PUs) for them to be executed. Actions express computational needs associated to messages and can require the use of resources belonging to the external environment. Actions are executed in parallel, depending on the availability of PUs. In general an action, after its submission by an actor, can run to completion or it can be suspended/resumed or aborted.

An action is a black box with a list of input parameters and a list of output parameters. Actions have no visibility to the internal data variables of the submitter actor. As a consequence, no mutual exclusion mechanism is required and no interference can occur from the action parallel execution schema. When an action terminates, it can inform the submitter by an action completion message. The submitter can then access the output parameter list to get any result computed by the action.

Actions can be reified in different ways. Simulated actions consist of pure time consuming activities whose aim is to advance the simulated time. Real or effective actions have a concrete instruction body (algorithm) whose execution advances the real time. Pseudo real actions increases the real time but have no concrete algorithm to execute. They can be useful for *preliminary real-time execution* of a given model (see later in this paper) which is a key to check how the overhead introduced by message exchanges and message processing affect the system timing constraints.

As a further refinement, action execution can be atomic or it can be preempted. In addition, an action can express an imprecise computation which after a time threshold delivers a first result whose accuracy can be improved would more time be available, or it can be returned and the action execution interrupted.

The various notions of actions are handled by the corresponding *action schedulers* provided by the control machine. An action scheduler manages local processing units and stores actions which find no available PU in pending action queues, waiting for some specific or unspecific PU to be ready to accept a new action execution. A PU can be a physical core or it can be realized by a Java thread, or it can be a fake object in the case of simulated actions. The use of preemptive actions/PUs were used in (Cicarelli&Nigro,2016a) to enable schedulability analysis of real-time systems.

A key factor of the actor control framework is *model continuity*, that is transitioning a same model from property analysis to real time execution. Model continuity mainly depends on actions. Moving from simulation to real execution requires changing the control machine, the time notion and the nature of actions which are switched

from simulated actions to real actions and associated action schedulers. All the remaining part of the model, and particularly actor behaviors and message passing, remains exactly the same during the transition.

Control framework in Java/JADE

Fig. 3 recapitulates some of the fundamental classes of the control framework. Actors and control machines are mapped on JADE agents. Each control machine owns an action scheduler which administers a set of processing units. A control machine receives the submitted actions and forwards them to the action scheduler. Actions and messages are embodied as serialized objects within *ACLMessages* when exchanged between actors and control machines.

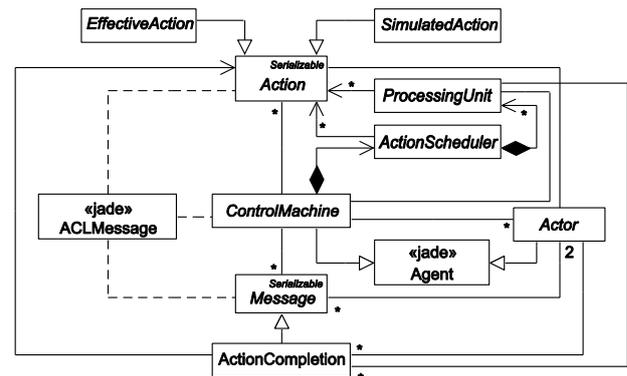


Figure 3 - Framework basic classes

The prototyped control machines are partitioned into three groups (see Fig. 4): (i) the untimed control machines, (ii) the time-aware control machines which operate in a sequential setting and (iii) the time-aware control machines which operate in a distributed context. A time server is required by the latter group in order to ensure a coherent time evolution among all the participating control machines (more details in (Cicarelli&Nigro, 2016a-b)).

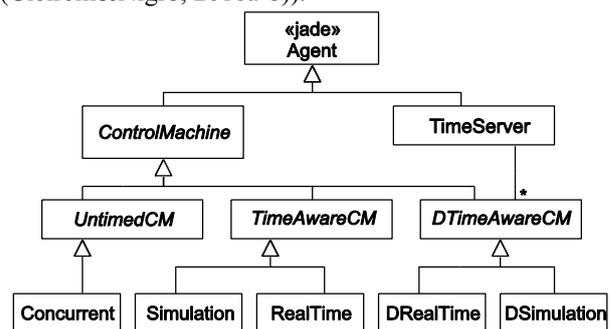


Figure 4 - Class hierarchy of control machines

Simulation, along with its parallel/distributed counterpart *DSimulation*, implements a classical discrete-event simulation schema. Messages are tagged with an absolute timestamp and are buffered into a time ranked queue where the head message holds the (or is one message with) minimum timestamp. The control machine can work with simulated actions. A simulated action carries the time duration of the associated activity. At its

submission, a simulated action is assigned to an exploitable PU which in this case simply means that an action completion message is scheduled with timestamp $now + duration$.

RealTime is a time-sensitive control machine using a real time notion built on top of the *System.currentTimeMillis()* *System* Java service. Only effective actions can be used. Messages have a timestamp and must be dispatched as soon as the current time exceeds their firing time. *RealTime* uses a configurable time tolerance *EPS*, so that a time-constrained message which should occur at absolute time t , is considered to be still in time if the current time is less than or equal to $t + EPS$. *RealTime* is useful for non-hard real-time applications. The *DRealTime* control machine replaces *RealTime* in the parallel/distributed context.

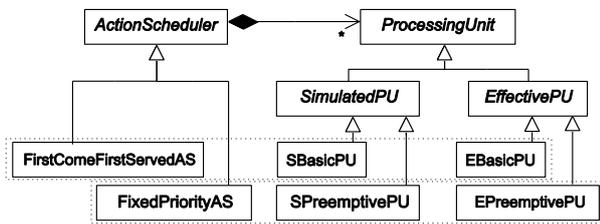


Figure 5 - Class hierarchies of action schedulers and PUs

As shown in Fig. 5, corresponding classes exist for action schedulers and PUs, which work together. Schedulers immediately put into execution a submitted action on an idle exploitable PU (if there are any), otherwise, different scheduling strategies can be adopted. In the case no such idle PUs exist, the scheduler *FirstComeFirstServedAS* organizes actions in a pending list. Actions will be executed according to their arrival time. The *FixedPriorityAS* uses instead an action priority to keep ordered the pending list. In this case, action execution is priority driven and preemptive.

For simulation purposes, the use of classes which are heirs of *SimulatedPU* is required. They are passive objects without inner threads. During real-time execution, heirs of *EffectivePU* should instead be used. They are thread-based objects able to execute effective actions (Cicirelli&Nigro, 2016a-b).

AN AGENT-BASED STPN TOOL

A class diagram of an *sTPN* tool built on top of the actor-based control framework is summarized in Fig. 6.

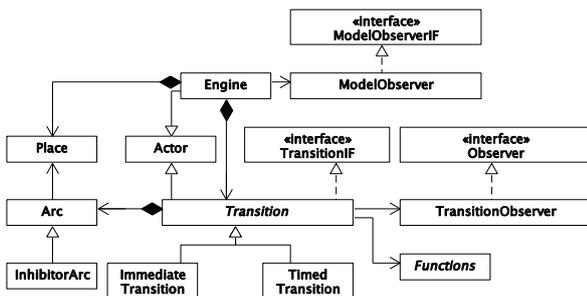


Figure 6 - Main classes in the *sTPN* tool

A model is feed through an XML file which is parsed into an internal representation consisting of a multi-agent system (MAS). Agents, i.e., actors, are associated with transitions which interact with the *Engine* (another agent) through messages and method calls. Each transition refers its input/output arcs which are linked to their input/output places. Arcs and places are realized as POJOs and provide their services by method invocations.

For analysis purposes, an *sTPN* model is simulated by associating to the generated MAS (single LP) the *Simulation* control machine (see Fig. 4) along with basic simulated actions paired with a first come first served scheduler (Fig. 5). Evolution of the MAS is triggered by transition firings and is controlled by the *Engine* which repeats a basic loop of simulation steps. At each step, the candidate set of enabled transitions is recomputed. The *Engine* owns the collection of transitions and the collection of places (marking) of the model. Transition enabling is checked by a method call on the transition agents.

A critical issue concerns the atomic firing process of a transition. Towards this the *Engine* separately executes the withdraw and the deposit phases of a transition firing. Each phase is immediately followed by a *retract* of disabled transitions which are removed from the candidate set and their firing messages invalidated in the simulation calendar. The two phases are required for correctly handling effective conflicts among transitions. An enabled transition can loss its enabling status in the intermediate marking following the withdraw phase or in the final marking reached after the deposit phase. Purposely, withdraw and deposit operations are realized by method calls (defined in the interface *TransitionIF* in Fig. 6) on the transition actor (recall that the computational status of an actor system is frozen between two consecutive message dispatches; therefore, method invocations, also with side-effects, are compliant with the actor lifecycle).

Immediate transitions, ranked according to their weights, are fired one at a time directly by the *Engine*. Timed transitions are instead fired by messages and actions. In particular, all the enabled timed transitions at a simulation step receive a *StartFiring* message from the *Engine* whose processing implies the next sample of the associated probability distribution function is obtained. The sample is passed to a submitted action which simulates the transition firing by scheduling the message completion message at the absolute time of $now + sample$. When the transition receives the *ActionExecutedMessage*, it informs the *Engine* about commitment of the transition firing through an *EndFiring* message. The *Engine* then executes the two phases *withdraw + retract*, *deposit + retract* on the transition. After that, the next step (iteration) of the *Engine* is started.

For property checking, a model evolution (see Fig. 6) is watched by suitable observers which collect statistical information about transitions or the entire model.

Since the *E*, *Uw* and *Ud* functions are model specific, the adopted solution consists in specifying, in the model

XML, the name of a Java class which provides an implementation of the above functions as methods. Such a class, subtype of the *Functions* abstract class (see Fig. 6), is then dynamically loaded, instantiated and exploited by transitions.

For preliminary execution of an *sTPN* model, the corresponding MAS is plugged with the *RealTime* control machine, and works with pure time consuming effective actions and the *FirstComeFirstServedAS* scheduler. All of this ensures the “effective actions” behave as in simulation but now advance the real-time.

A CASE STUDY USING MODEL CONTINUITY

The Fisher’s mutual exclusion protocol for N processes, having identifiers $1, 2, \dots, N$, competing for the access to some shared resource, was used as a case study for validating the obtained *sTPN* tool. The protocol is an example of a time-dependent mutual exclusion algorithm (Lynch&Shavit, 1992). Although the algorithm has been analyzed qualitatively by model checking, e.g., in the context of the UPPAAL toolbox (Behrmann et al., 2004), here it is used for quantitative evaluation using simulation and the results compared with those described in (Paolieri et al., 2016) which are based on probabilistic model checking and numerical approach.

The protocol assumes that basic read/write memory operations are atomic. A single global communication variable id is used, which stores the identifier of the process trying to enter its critical section, or it defaults to 0. Every process can try the protocol when $id = 0$. In this case the process executes the time-consuming operation $id \leftarrow i$. Since more processes can attempt the same operation simultaneously, it is required for a trying process to wait for a time (say it W^+) greater than the writing time W . After W^+ time units are elapsed, the process reads again id . In the case $id \neq i$ the process has to retract and to wait for the id to become again 0. If instead $id = i$ then the process can enter its critical section. At the exit from the critical section, the process sets id to 0.

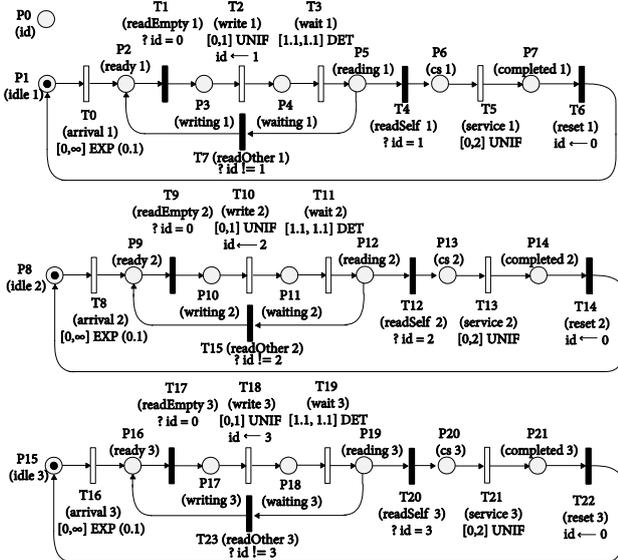


Figure 7 – A model of Fisher’s protocol (Paolieri et al., 2016)

The model in Fig. 7 with $N = 3$ processes, derived from (Paolieri et al., 2016), was used for the experiments. Immediate transitions are shown as black bars, whereas timed transition are depicted as white bars. The non critical section of each process is modelled by an *arrival* timed transition with interval $[0, \infty]$ and with an exponential pdf (*EXP*) with rate 0.1. W is supposed to be uniformly distributed (*UNIF*) within the interval $[0, 1]$, and W^+ is set to 1.1 time units. Other details of the model should be self-explanatory.

Property analysis

The Fisher’s *sTPN* model was studied in (Paolieri et al., 2016) using a probabilistic temporal logic built around an *interval until operator*:

$$\varphi U^{[\alpha, \beta]} \psi$$

which captures the event that a marking of the *sTPN* model is reached which satisfies a predicate ψ at some time in the interval $[\alpha, \beta]$ without violating a safety predicate φ . Of such event can be of interest finding the occurrence probability P , or bounding such probability against a given threshold value: $P \sim_p$ where $\sim \in \{<, >\}$. Predicate states are based on net markings. As usual, predicates and atomic propositions can be combined with boolean operators to form more complex formulas, but nesting of interval until operators is not allowed.

The interval until operator naturally can be used to assess transient behavior of a net model.

The following properties were studied using the Fisher’s protocol model upon the developed *sTPN* tool: (a) mutual exclusion (*safety*), i.e., no more than one process can enter its critical section at a time; (b) absence of starvation (*bounded liveness*), that is a trying process eventually enters its critical section. The latter property relates to estimating the overtaking factor, i.e., the maximum number of by-passes of other processes with respect to a waiting process, or equivalently to estimating the maximum waiting time of a trying process before entering its critical section; (c) some other examples of specific bounded liveness properties. In each case a proper decoration of the model observer was used. In the following, for simplicity, the notation, e.g., cs_1 is used instead of $m[cs_1]$.

Mutual exclusion

Mutual exclusion was checked by performing some simulation runs with $t_{End} = 3.5 \times 10^5$ time units, and by observing that the event

$$true U^{[0, t_{End}]} (cs_1 = 1 + cs_2 = 1 + cs_3 = 1 > 1)$$

has a 0 probability of occurrence. The model observer object was decorated to watch marking of cs_1 , cs_2 and cs_3 places. In no case it was found more than one process is in its critical section. As part of this assessment it was also checked that effectively it can happen that csi for any i assumes the value 1.

Absence of starvation

It was estimated the probability that a process can be affected by a certain number of by-passes (overtaking) from other competing processes. As one can see from Fig. 8, each process seems to suffer for no more than 6 by-passes. 5 simulations with $t_{End} = 3.5 \times 10^5$ time units were used to collect data behind Fig. 8 and Fig. 9.

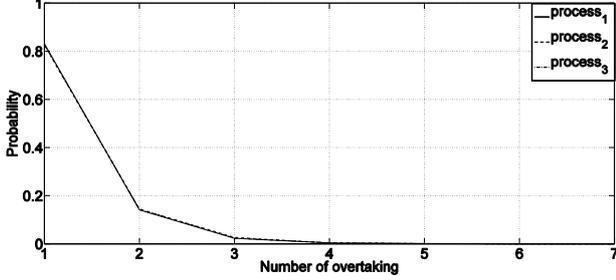


Figure 8 - Occurrence probability vs. number of by-passes

Another way to check the starvation-free behavior was estimating the worst case waiting time of a trying process. Results are shown in Fig. 9.

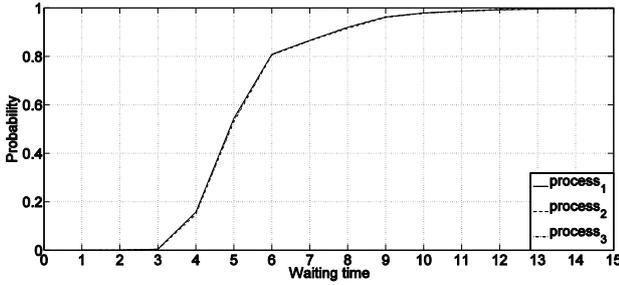


Figure 9 - Trying process waiting time

Examples of bounded liveness properties

As an example of a particular bounded liveness property it was measured the occurrence probability of the following event:

$$true U^{[0,\beta]}(cs1 = 1) (*)$$

for various values of β , starting separately from each of the following markings which describe possible execution states of the three processes:

$$\begin{aligned} m_A &\equiv ready_1 idle_2 idle_3 \\ m_B &\equiv id = 3 ready_1 idle_2 waiting_3 \\ m_C &\equiv id = 3 ready_1 waiting_2 waiting_3 \end{aligned}$$

The property addresses specifically a deadline requirement upon the delay $process_1$ experiments before entering its critical section.

A batch of simulation runs were carried out, terminating each of them as soon as the watched event occurs (that is the given number of by-passes happens). The proportion of the runs which satisfy the event divided by the total number of runs was then evaluated.

The number of required runs was empirically determined by watching the probability value which

almost stabilizes. 100 runs were used for building each curve in Fig. 10.

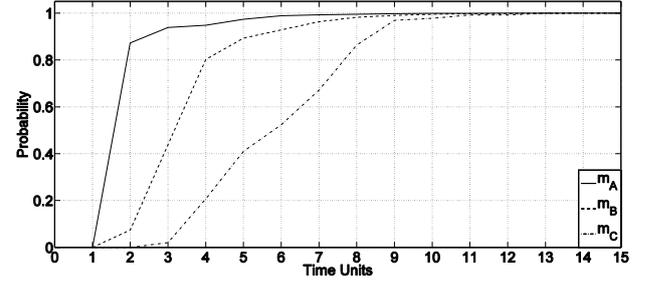


Figure 10 - Occurrence probability of the event (*) vs. time

For example, in the m_A scenario, as expected, the delay of $process_1$ can be short (best cases) with a probability of about 90% but in the worst case (probability 1) it amounts to known maximum waiting time (Fig. 10).

Another particular property concerned an evaluation of the occurrence probability of the following event:

$$!cs_1 U^{[0,\beta]}(completed_2 \vee completed_3) \wedge (ready_2 \vee ready_3) (**)$$

for various values of β , starting from the marking

$$ready_1 idle_2 idle_3$$

and separately for three different service time distributions: $UNIF[0,2]$, $UNIF[0,4]$, $UNIF[2,4]$. The event amounts to asking the following check: in the hypothesis that $process_1$ is not in its critical section, what is the worst case time (β) for each of the remaining processes so as to be ready to try or be capable of having completed an access to shared data? Respectively 235, 195 and 220 runs were used for generating the three curves in Fig. 11.

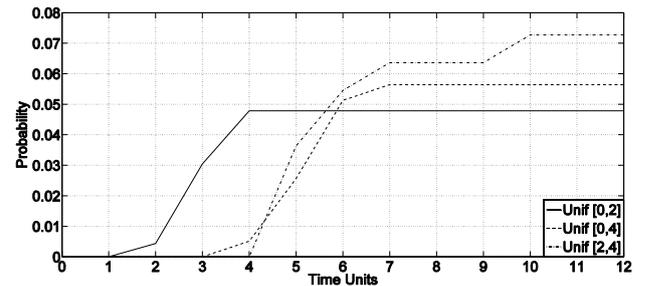


Figure 11 - Occurrence probability of event (**) vs. time

It is worth noting that the results portrayed in Fig. 10 and Fig. 11 are very close to the results reported in (Paolieri et al., 2016) for the same checked events.

Preliminary real-time execution

After property analysis, the sTPN Fisher's protocol was re-checked by executing it in real-time but with pure time consuming actions instead of effective instructions of a concrete programmed version of the process bodies.

Such *preliminary execution* is very important in the practical case for observing the overhead introduced by

scheduling and message exchanges on the fulfillment of model timing constrains. Configuring the model for preliminary execution only required: (a) interpreting the time unit as *1 sec*; (b) changing the control machine from *Simulation* to *RealTime* and (c) using pure time consuming effective actions with the *FirstComeFirstServedAS* scheduler. No changes were introduced in the model. The Fisher's protocol was then executed with a time tolerance of $EPS = 200\text{ ms}$.

Basic properties of the mutual exclusion algorithm were watched during the execution and the time deviations, i.e., the latency with which messages and actions are actually executed with respect to their due time, measured.

Worst case results of 4 runs each of 7 hours of wall clock time are collected in the histogram of Fig. 11. As one can see, in almost all the cases, the time deviation is virtually 0 ms. The most frequent non zero deviation is 16 ms. The worst case deviation was found to be 155 ms which occurred just once at the execution start, i.e., at model bootstrapping.

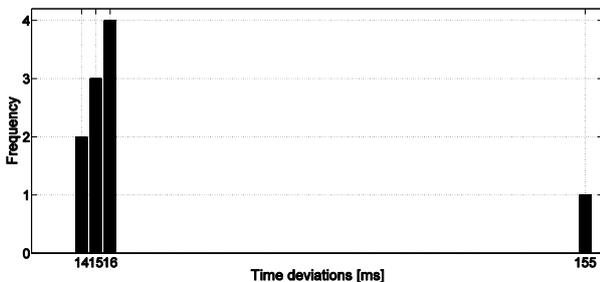


Figure 11 - Observed time deviations

During the whole real-time experiment, no more than one process was found in its critical section. In addition, in Fig. 12 is portrayed an histogram of registered overtaking factor and its occurrence probability of trying processes.

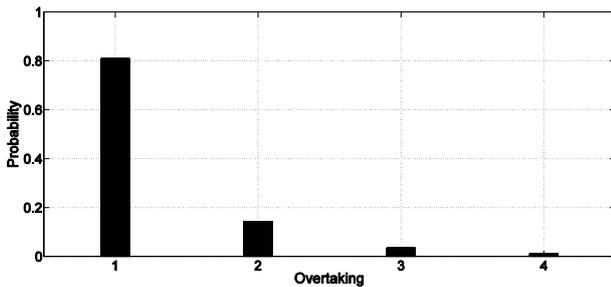


Figure 12 - Observed overtaking

All the experiments were carried out on a Win 7 workstation, 12GB, Intel Core i7, 3.50GHz, 4 cores, without an active Internet connection.

CONCLUSIONS

Although potentially less accurate of the probabilistic model checking approach based on numerical solutions proposed in (Paolieri et al., 2016), the agent-based simulation tool for Stochastic Time Petri Nets (sTPN) described in this paper proves effective in the practical case, as demonstrated by the reported case study.

A key factor of the approach is *model continuity*, i.e. the same model can be used for property checking by simulation and for real-time execution. The tool features derive from the adopted underlying agent-based control-centric framework (Cicirelli&Nigro, 2016a-b).

The proposed approach provides the abstraction mechanisms and the execution concerns suited, e.g., for modelling, analysis and execution of time-constrained workflow systems. The real-time preliminary execution, in particular, directly corresponds to workflow enactment.

Prosecution of the research work is geared at:

- optimizing the implementation of the sTPN tool;
- applying the approach to modelling, analysis and enactment of time-constrained workflow systems (Gonzales del Foyo&Silva, 2008);
- supporting a probabilistic temporal logic (Younes et al., 2006)(Paolieri et al., 2016)(David et al., 2015) for the expression of quantitative properties to check on an sTPN model, and automating the determination of the required simulation runs;
- extending the tool toward parallel/distributed simulation of large models.

REFERENCES

- Behrmann, G., A. David, K.G. Larsen (2004). A tutorial on UPPAAL. In: *Formal Methods for the Design of Real-Time Systems*, M. Bernardo and F. Corradini Eds., Lecture Notes in Computer Science, Vol. 3185, Springer-Verlag, pp. 200-236.
- Bellifemine, F., G. Caire, D. Greenwood (2007). *Developing multi-agent systems with JADE*. John Wiley & Sons.
- Bucci, G., L. Carnevali, L. Ridi, E. Vicario (2010). ORIS: a tool for modeling, verification and evaluation of real-time systems. *Int. J. on Software Tools for Technology Transfer*, Springer, vol. 12, pp. 391-403.
- Cicirelli, F., C. Nigro, L. Nigro (2015). Qualitative and quantitative evaluation of stochastic time Petri nets. Proc. of *2nd Int. Workshop on Cyber-Physical Systems (IWCPs'15)*, Lodz, Poland, pp. 775-784.
- Cicirelli, F., L. Nigro (2016a). Control aspects in multi-agent systems. In *Intelligent Agents in Data Intensive Computing*, Springer, Studies in Big Data, Kolodziej J., Correia L., Manuel Molina J. (Eds.), pp. 27-50.
- Cicirelli, F., L. Nigro (2016b). Control centric framework for model continuity in time-dependent multi-agent systems. *Concurrency and Computation: Practice and Experience*, Wiley, to appear.
- David, A., K.G. Larsen, A. Legay, M. Mikucionis, D.B. Poulsen (2015). UPPAAL SMS Tutorial, *Int. J. on Software Tools for Technology Transfer*, Springer, 17:1-19, 06.01.2015, DOI 10.1007/s10009-014-0361-y, 2015
- Gonzalez del Foyo, P.M., J.R. Silva (2008). Using Time Petri Nets for modelling and verification of timed constrained workflow systems. In *ABCM Symposium Series in Mechatronics*, vol. 3, pp.471-478.
- Lynch, N., N. Shavit (1992). Timing-based mutual exclusion. In *IEEE Real-time Systems Symp.*, pp. 2-11.
- Paolieri, M., A. Horváth, E. Vicario (2016). Probabilistic model checking of regenerative concurrent systems. *IEEE Trans. Soft. Eng.*, to appear (available on-line).
- Younes, H.L.S., M. Kwiatkowska, G. Normaln, D. Parker (2006). Numerical vs. statistical probabilistic model checking. *Int. J. on Software Tools for Technology Transfer*, vol. 8, no. 3, pp. 216-228.

FRONTIER BASED MULTI ROBOT AREA EXPLORATION USING PRIORITIZED ROUTING

Rahul Sharma K.

Daniel Honc

František Dušek

Department of Process control
Faculty of Electrical Engineering and Informatics,
University of Pardubice, Czech Republic
E-mail: rahul.sharma@student.upce.cz
{daniel.honc, frantisek.dusek}@upce.cz

Gireesh Kumar T

TIFAC CORE In Cyber Security

Amrita School of Engineering

Amrita Vishwa Vidyapeetham University

Coimbatore, India

E-mail: t_gireeshkumar@cb.amrita.edu

KEYWORDS

Multi robot, Area exploration, Path Planning, Frontier based, Optimization, Agent based simulation.

ABSTRACT

The paper deals with multi-robot centralized autonomous area exploration of unknown environment with static obstacles. A simple reasoning algorithm based on pre-assignment of routing priority is proposed. The algorithm tracks the frontiers and assigns the robots to the frontiers when the robots fall into a trap situation. The algorithm is simulated with various multi-robotic configurations in different environments and compared with performance indices in the MATLAB simulation environment.

INTRODUCTION

Area exploration is one of the fundamental problems of autonomous robotics. The main goal of any exploration algorithm is to gain as much new information as possible of the unknown environment within the bounded time. Autonomous area exploration algorithms find applications in space robotics, military operations, disaster management, sensor deployment etc. Area exploration deals with exploring through all unknown areas and creating a map of the environment. Most of the area exploration keeps a map of the environment and updates when an unknown region is explored.

Yamauchi pioneered the research in the frontier-based area exploration (Yamauchi 1997). A frontier is a boundary that separates known (explored) regions from unknown regions. By moving towards frontiers, robots can focus their motion on discovery of new regions. The frontier based area exploration is extended to multi-robot system (MRS) in (Yamauchi 1998). (Yan et al. 2013) presented a systematic survey and analysis of coordination of multiple mobile robot systems. A comparative study of area exploration algorithms can be seen in (Dayanand et al. 2013).

An important task in an area exploration algorithm is “how the robots choose which cell is to be explored next”. The aim of any frontier based algorithm is to explore all the frontiers in shortest time possible. In

multi-robot exploration, the robots coordinate each other and decide which robot will explore which frontier, based on a coordination / routing policy. Various coordination policies (Burgard et al. 2000, Burgard et al. 2005, Ma et al. 2006 and Wang et al. 2011) have been proposed in the past. In this paper, we propose a simple reasoning based routing policy determined by a pre-assigned priority. A central agent (CA) coordinates with a group of multi robot agents (MRAs) to explore an unknown environment with only static obstacles. The MRAs, with the command from CA, move from one cell to another cell, sense the environment and communicate the information to the CA. The CA will make the routing decision based on the state of the adjacent cells and the pre-assigned priority of the corresponding MRA. The CA will also keep a list of frontiers and assigns MRAs, once a MRA falls into a “trap” situation. A MRA is said to be in a *trap* situation if all the adjacent cells are either explored and/or occupied with obstacles. The route to the frontier is the shortest path found by executing the A* algorithm (Hart et al. 1968). The algorithm terminates when all the cells are explored and no more frontier cells exist. The key advantage of prioritised routing is that, the predictability of the MRAs location is improved. This will help in a decentralised distributed coordination policy, where MRAs collectively make the routing decision as MRAs can predict each other’s location with the help of a pre-assigned priority of routing.

FRAME WORK FOR AREA EXPLORATION

Environment

The unknown environment is considered as grids with cells of same dimensions. If any of the cells is occupied with an obstacle, the whole cell would be considered as occupied.

Mobile robot agent

A general scheme of MRA is shown in figure 1. The following are functions of MRA

- Receive motion commands (e.g. Move Forward, Right, Left or Backward).
- Execute the commands by a motion control algorithm, for e.g. PID control with wheel

encoders and motors as actuators to make the mobile robot move from one cell to another.

- Sense the surrounding cells using (short range) distance sensors, like infra-red or ultrasonic sensors.
- Send the information about the environment to the CA.

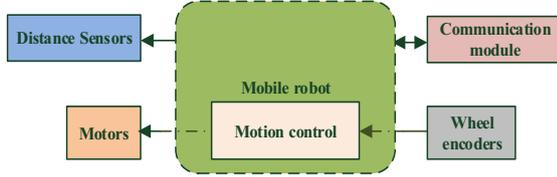


Figure 1: Scheme of Mobile Robot Agent

The mobile robots (for e.g. Sharma 2013) are assumed to be

- Point-sized and occupy only one grid at a time.
- Assigned with a fixed priority in path planning.
- Move at an even speed, and its status can be switched - between moving with a fixed speed and halting.
- In any of the following orientations - North, East, South, West.
- Able to move in four directions (**F**, **R**, **L**, **B**). Each robot can move from one cell to four adjacent neighbouring cells (other than the diagonal cell) in one step.

Central Agent

The central agent can be a stationary computer or a mobile robot with sufficient computational power and a communication module. The following are the functions of the CA.

- Receive commands from the MRAs
- Update the map
- Frontier detection
- Trap detection and run shortest path planning algorithm
- Routing policy
- Send commands to the MRAs

Problem description

- To explore an unknown environment using multi-robots in the shortest time possible.
- An OPEN cell shouldn't be explored more than once and should follow the shortest path to the frontier.
- While exploring, no collision with the robots at any point of time.
- To detect any robot in a trap and assign them to move to any frontier.
- To keep track of frontiers and assign the robot to that frontier, in the shortest possible path, by executing a shortest path planning algorithm.

Assumptions

- The robots can communicate with the central computer or agent without information (packet) loss or any time delay and vice versa.
- The environment is considered as grids with cells as same length and width ($L \times L$).
- A grid based environment is assumed (if any part of the cell is occupied with obstacle the whole cell is considered as occupied).
- The maximum sensing range of MRAs are just one cell length (L).

Terminology

Each *cell* is assumed in any one of the following states:

- OPEN – the cell is not explored and no obstacle is present (detected by the robot, but not visited)
- OCCUPIED– the cell is explored and obstacle is present
- UNKNOWN – the cell is not explored (neither visited nor detected by the robot)
- CLOSED – the cell is explored and no obstacle is present

A robot is said to be in a *trap* situation, when there are no possible moves to any of the unexplored adjacent (OPEN) cells. In other words, if the adjacent cells are either CLOSED or OCCUPIED. A robot is said to be in *on_command* state when it is in pursuit of exploring a frontier. The attributes used in the simulation are shown in Table 1 and 2.

Table 1: Attributes of MRAs

Attributes, $r(\bullet)$	Meaning	Type
r_num	ID of robot	int
$r_priority$	pre-assigned priority of robot	int 2D array
r_cxy	current position	int array
r_nxy	next position	int array
$r_corient$	current orientation	int
$r_norient$	next orientation	int
r_trap	if trap-1 ,else - 0	boolean
r_status	status of robot	int
r_on_comm and $route$	route to the frontier	int 2D array

A robot will be in any of the following states,

$$r_status = \begin{cases} 1, & \text{if robot is in } trap \text{ situation} \\ 2, & \text{if robot is } on_command \text{ state} \\ 0, & \text{else} \end{cases}$$

The fr_list represents a list of all explored frontiers and n_r , the number of robots. The status of the frontier is assumed as,

$$fr_closed = \begin{cases} 1, & \text{if a robot already assigned} \\ 0, & \text{if frontier is detected, but not assigned} \end{cases}$$

Table 2: Attributes of the Frontiers

Attributes, $fr(\bullet)$	Meaning	Type
fr_number	cell number of the frontier	int array
fr_length	path length to the frontier	int
$fr_waiting$	number of iterations waiting from detection of frontier	int
fr_route	route to reach the frontier from current location of the robot	int 2D array
fr_closed	status of frontier	boolean
fr_robo	ID of robot assigned to explore the frontier	int

FRONTIER BASED AREA EXPLORATION ALGORITHM BY PRIOROTIZED ROUTING

The algorithm terminates when there are no more frontiers or OPEN cells present. The following are the steps involved in the algorithm.

Priority assignment strategy

The MRAs will be pre-assigned with a fixed priority, which will not change during the course of the area exploration iterations. A sample assignment strategy is mentioned in Table 3 for three MRAs. The basic idea of this assignment is that, when a robot encounters a junction, for example both right and left side are unexplored and OPEN, the robot#1 chooses the left turn and robot#2 will choose right turn. If the 1st priority movement is not possible (already explored or OCCUPIED), then the robot chooses the 2nd priority and if that is also explored then, the robot chooses 3rd priority and so on. The fourth priority is assumed as a backward motion, as it is assumed that area exploration proceeds forward.

Table 3: A Sample Priority Assignment Strategy for 3 Robot MRS

Priority ID	1 st	2 nd	3 rd	4 th
#1	L	F	R	B
#2	R	F	F	B
#3	F	R	L	B

New frontier detection

The master robot maintains a list of frontiers to be explored. Whenever a new frontier is discovered, the master robot will add the details of the frontier into the database. A frontier is detected when a robot senses unexplored cells (OPEN) in more than one side. The robot will move to one of the OPEN cells based on the priority assigned to it. The other unexplored cell information will be stored as frontiers and these frontiers will be explored once a MRA enters the *trap* situation.

Table 4: Cell Increments when Robot Moves from One Cell to Another with Respect to Orientation

Motion Orientation	F	L	R	B
N	[-1, 0]	[0, -1]	[0, 1]	[1, 0]
S	[1, 0]	[0, 1]	[0, -1]	[-1, 0]
W	[0, -1]	[1, 0]	[-1, 0]	[0, 1]
E	[0, 1]	[-1, 0]	[1, 0]	[0, -1]

Receiving the commands from MRAs and updating the map

At every time instance, the MRAs will send information (F, L, R or B) about the adjacent cells with their ID as a header. For e.g. an information from an MRA saying #1FR means, the robot#1 found the cell just in front of the current cell and the cell to right side is OCCUPIED (i.e. sensors detects obstacles). The first step is to initialize the map (grid) with UNKNOWN status. The map is updated with the information from MRAs as shown below.

Algorithm 1 : Updating the map

```

1: for  $i=1:n_r$ 
2:   Get the coordinates of the sensed cells by robot
   using  $r(i).cxy$  and Table 4(motion,  $r(i).c\_orient$ )
3:   Mark grid( $x, y$ ) as OCCUPIED
4: end for

```

Routing policy, Trap detection and Frontier detection

Algorithm 2 : Routing policy, trap detection, frontier detection

```

1: for  $j=1:n_r$ 
2:   if  $r(j).status$  is not in trap or on command
3:     for  $i=1:r(j).prio$ 
4:       get the coordinates ( $x, y$ )
5:       if grid( $x, y$ ) is OPEN
6:         if next position is not found
7:           update  $r(j).nxy$  (using Table 4) and
            $r(j).norient$  as  $i$ 
8:         end if
9:       else
10:        if ( $x, y$ ) not in  $fr\_list$ 
11:          Mark it as  $fr$  and update  $fr\_list$ 
12:        end if
13:      end if
14:    end for
15:    if no next position or frontier detected
16:      mark robot  $r(j).trap=1$  and  $r(j).status=1$ 
17:    end if
18:  end if
19: end for
20: end for

```

The CA keeps a list of all frontiers detected and explored. The CA checks the adjacent cells of the current positions of the MRAs and makes the routing policy. With respect to the priority assigned for each robot, the CA checks the

adjacent OPEN cells in the order of priority and assigns the first encountered cells as next exploration cell, and all the following cells as frontiers. If the CA couldn't find any OPEN unexplored cell, the robot will be marked as in the *trap* situation. Algorithm 2 shows the steps involved.

Robot assignment to frontiers in case of trap situation

A robot is said to be in a *trap* situation, when there are no possible moves to any of the unexplored adjacent cells. This condition must be carefully studied; otherwise there are chances of the mobile robot getting in to an infinite loop. When the CA detects any robot in a *trap* situation, it looks for any frontier cells yet to be explored. If any frontier cell exists, the CA runs A* shortest path planning algorithm to find shortest distance to the frontier with source node as a robot's current position and destination node as a frontier cell. The algorithm returns the shortest path, if it exists. Algorithm 3 describes the steps involved.

Algorithm 3 : In case of trap situation

```

1:   if any of the robot in trap
2:     Get the ID  $j$  of the robot
3:     if  $fr\_list$  is not empty
4:       get the coordinates of frontier  $k$ 
5:       find the shortest path  $route$  by A* algorithm
6:       if path exists
7:         remove  $k$  from  $fr\_list$ 
8:         Mark  $r(j).status=2, r(j).trap=0$  and
            $r(j).on\_command\_route=route$ 
9:       end if
10:    end if
11:  end if

```

Send the commands to MRAs

The CA will send the commands (**F**, **R**, **L** or **B**) to MRAs with the header as MRAs ID, based on the routing policy as explained earlier. In case of *on_command* status the CA will send the command ($r_on_command_route$) at every time instant, one by one, calculated by executing the A* algorithm.

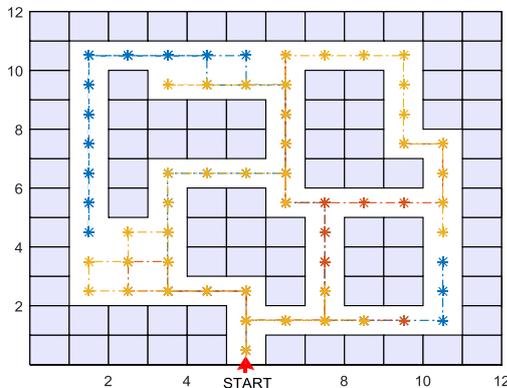


Figure 2: Stairway Exploration with Three Robots (Red, blue, orange – robot#1, #2 and #3 respectively)

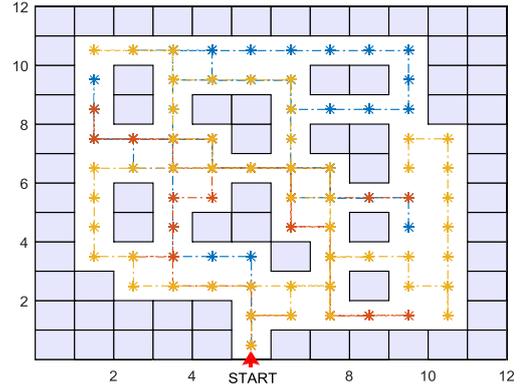


Figure 3: Unstructured Environment Exploration with Three Robots (Red, blue, orange – robot#1, #2 and #3 respectively)

SIMULATION RESULTS

The simulation studies were performed in MATLAB with two different environments - stairway and unstructured environment. The initial orientation of the robot were assumed to be the robot facing the North direction. Figure 2 and 3 shows the performance of the algorithm using three MRAs in a stairway and unstructured environment.

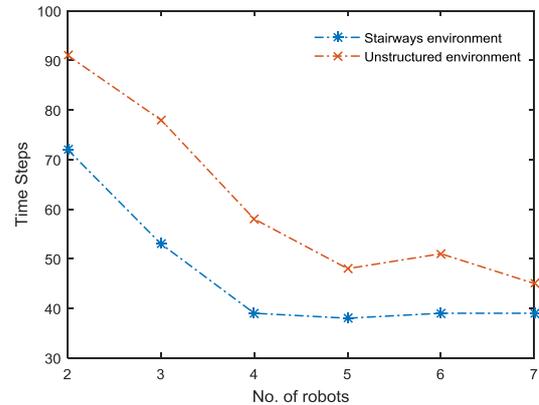


Figure 4: Time Steps Vs Number of Robots

Experiments were conducted with different robot configurations and different priorities were assigned to the MRAs. Figure 4 shows the simulation time steps taken for different robot configurations. The efficacy of different configurations can be compared with the performance indices.

Performance Indices (PI)

Five PI's were chosen to measure the performance of the different robot's configuration using the proposed algorithm.

Iterations: The number of time steps taken to explore the whole unknown environment.

Frontiers explored (n_{fr}): The number of frontiers detected by the CA and explored.

Average path length (L_r): The ratio of total number of cells visited by the robots (S_r) to total number of robots (n_r).

Exploration efficiency index (η_e): It is the measure of the efficiency of the exploration algorithm that how efficiently (without re-visiting the same cells again and again) the robots are able to explore the environment. In other words, it is a scale of how many cells the robot revisited during the course of exploration of the frontier cells. A high value represents less efficient and vice versa. It is given by the following relation,

$$\eta_e = \frac{L_r \times n_r - S_c}{S_c} \times 100\%$$

Where S_c represents total number of cells without obstacle.

Coverage: It is the ratio of the number of explored cells to the total number of cells without obstacle (S_c). The algorithm terminates at 100% coverage.

Table 5 and 6 shows the performance indices of the area exploration algorithm in the stairway (figure 2) and unstructured environment (figure 3) respectively. Figure 4 depicts the iterations vs the number of robots. It can be seen that, the four robots perform significantly better than 2 or 3 robots, while there is not much gain in having more than four robots.

Table 5: Performance Indices of Area Exploration in Stairway

PI \ n_r	2	3	4	5	6	7
iterations	72	53	39	38	39	39
n_{fr}	19	27	30	19	32	34
L_r	72	50	34	32	31	28
η_e (%)	145	155	133	174	222	237

Table 6: Performance Indices of Area Exploration in Unstructured Environment

PI \ n_r	2	3	4	5	6	7
iterations	91	78	58	48	51	45
n_{fr}	31	34	33	34	40	34
L_r	91	74	57	39	34	35
η_e (%)	144	198	204	161	174	232

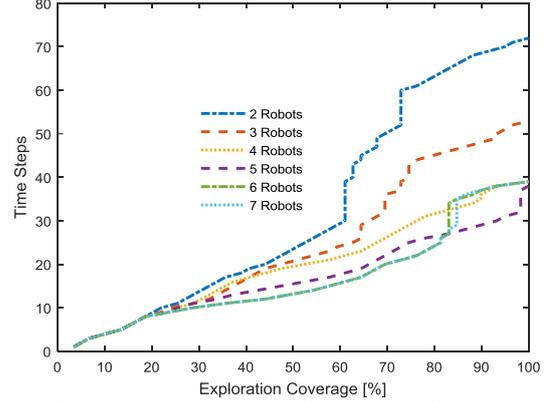


Figure 5: Coverage – Stairways Environment

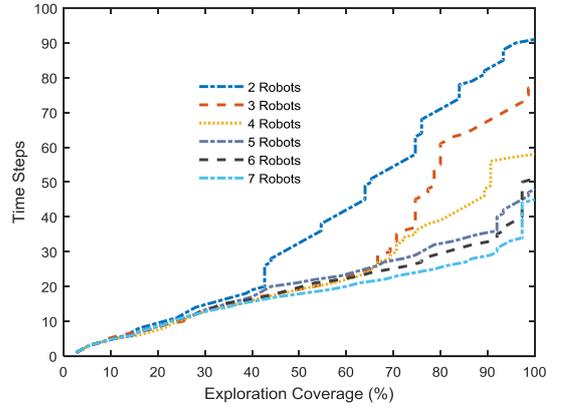


Figure 6: Coverage – Unstructured Environment

In order to say, which one is the best configuration (number of robots and pre-assigned priority) we need to have a tradeoff between the performance indices. For e.g. if only time taken is important, then the best option would be the maximum number of robots. If the PI's L_r and η_e are more important, then 6 robots for unstructured (figure 3) and 4 robots for the stairway (figure 2) would be the best option.

The main objective was to develop and simulate a frontier based area exploration algorithm with a pre-assigned priority. The performance of the algorithm heavily depends on two factors - number of robots and pre-assigned priority. The performance of the algorithm can be improved significantly, by introducing a cost function for the frontiers and the routing MRAs based on the cost function rather than waiting for the robot to fall into a *trap* situation.

CONCLUSION

The problem of area exploration was solved by a simple reasoning based frontier area exploration algorithm. The algorithm is based on centralized CA-MRA coordination. The CA makes the routing decision based on the status of MRAs and the frontiers yet to be explored. The routing policy is based on a pre-assigned priority. The main advantage of the proposed priority based approach, when

compared with random routing, is that the predictability of location of MRAs increases, and this will help in a more decentralized distributed exploration policy.

As a future research direction, we are looking forward,

- To increase the sensing range of the MRAs, which will have more practical implication.
- To develop a greedy approach based on a cost function for frontier exploration rather than waiting for any robot to enter a *trap* situation.
- To develop an advanced motion control algorithm for MRAs and to validate the algorithm by real experiments in the laboratory.

This research was supported by project SGS_2016_021, Mobile Robot Motion Control with Model Predictive Controller at FEL, University of Pardubice. This support is very gratefully acknowledged.

REFERENCES

- Burgard, W., Moors, M., Fox, D., Simmons, R. and Thrun, S., 2000. Collaborative multi-robot exploration. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on* (Vol. 1, pp. 476-481). IEEE.
- Burgard, W., Moors, M., Stachniss, C. and Schneider, F.E., 2005. Coordinated multi-robot exploration. *Robotics, IEEE Transactions on*, 21(3), pp.376-386.
- Dayanand, V., Sharma, R. and Kumar, G., 2013. Comparative Study of Algorithms for Frontier based Area Exploration and Slam for Mobile Robots. *International Journal of Computer Applications*, 77(8).
- Hart, P.E., Nilsson, N.J. and Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. *Systems Science and Cybernetics, IEEE Transactions on*, 4(2), pp.100-107.
- Ma, X., Meng, F., Li, Y., Chen, W. and Xi, Y., 2006, June. Multi-agent-based Auctions for Multi-robot Exploration. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on* (Vol. 2, pp. 9262-9266). IEEE.
- Sharma, R., 2013. Design and implementation of path planning algorithm for wheeled mobile robot in a known dynamic environment. *IJRET: International Journal of Research Engineering and Technology*, 2(06), pp.967-970.
- Yamauchi, B., 1997, July. A frontier-based approach for autonomous exploration. In *Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on* (pp. 146-151). IEEE.
- Yamauchi, B., 1998, May. Frontier-based exploration using multiple robots. In *Proceedings of the second international conference on Autonomous agents* (pp. 47-53). ACM.
- Yan, Z., Jouandeau, N. and Cherif, A.A., 2013. A survey and analysis of multi-robot coordination. *International Journal of Advanced Robotic Systems*, 10.
- Wang, Y., Liang, A. and Guan, H., 2011, April. Frontier-based multi-robot map exploration using particle swarm optimization. In *Swarm Intelligence (SIS), 2011 IEEE Symposium on* (pp. 1-6). IEEE.

AUTHOR BIOGRAPHIES



RAHUL SHARMA K., was born in Kochi, India and went to the Amrita University, where he studied electrical engineering and obtained his M.Tech degree in 2013. He is now doing his Ph.D. studies at the Department of process control, Faculty of Electrical and Informatics, University of Pardubice, Czech Republic.

e-mail: rahul.sharma@student.upce.cz



DANIEL HONC was born in Pardubice, Czech Republic and studied at the University of Pardubice in the field of Process Control and obtained his Ph.D. degree in 2002. He is head of the Department of Process

Control at the Faculty of Electrical Engineering and Informatics. e-mail: daniel.honc@upce.cz



FRANTIŠEK DUŠEK was born in Dačice, Czech Republic and studied at the Pardubice Faculty of Chemical Technology in the field of Automation and obtained his MSc. degree in 1980. He worked for the pulp and paper research institute IRAPA. Now he is the vice-dean of the Faculty of Electrical Engineering and Informatics. In 2001 he became an Associate Professor.

e-mail: frantisek.dusek@upce.cz



Gireesh Kumar T., was born in Wandoor, India, done Doctorate in Computer Science and Engineering, Anna University Chennai in 2011, domain of Cognitive Robotics. He is currently working as Associate Professor in Cyber Security, Amrita

School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India.

email: t_gireeshkumar@cb.amrita.edu

Effectively Operating Simulation

SIMULATION-BASED PERFORMANCE MEASUREMENT: ASSESSING THE PURCHASING PROCESS IN A PUBLIC UNIVERSITY

Pasquale Legato

Lidia Malizia

Rina Mary Mazza

Department of Informatics, Modeling, Electronics and System Engineering

University of Calabria

Via P. Bucci 42C

87036 Rende (CS), Italy

e-mail: {legato, lmalizia, rmazza}@dimes.unical.it

KEYWORDS

Performance modeling, business process, organization.

ABSTRACT

Performance measurement is becoming a must in the public sector in Italy, just as in other frontline economies. Public services have to be supplied to citizens under diminishing resources, but pursuing growing target levels as if they were operating in a competitive market. Discrete-event simulation is challenging as an effective methodology for a quantitative evaluation of different practices in non-profit organizations characterized by socio-technical environments guided by the central government's changing normative and often conflicting multiple stakeholders. This paper focuses on a scientific Department of an Italian University, after that a performance measurement and evaluation system has been adopted by the Board of Directors as required by recent laws aimed at increasing the level of accountability. A case study is described in which the "purchasing process" is analyzed by stochastic simulation in order to account for limited resources under various sources of uncertainty. Numerical results are presented to support possible managerial decisions towards improved efficiency, effectiveness and transparency in purchasing operations.

INTRODUCTION

Simulation is nowadays recognized as an effective methodology for computer-based (re)design of operations management systems under a dynamic stochastic environment (Shafer and Smunt 2004) and business processes (re)engineering in a socio-technical environment (Hlupic and de Vrede 2005; Gregoriades and Sutcliffe 2008). The development of specific business tailored paradigms (Melao and Pidd 2006) and friendly approaches (Robinson et al. 2014) to business process simulation (BPS) should encourage practitioners. The increasing dominant role of information systems and web services in the public sector call for revisiting BPS paradigms (Van der Aalst 2010) to also allow a useful integration of output analysis with spreadsheet-based tools for process

analysis (Saldivar et al. 2016). Some interesting applications have already demonstrated the benefit of BPS in the public sector (Haysa and Bebbington 2000; de Boer et al. 2003; Greasley 2006; Dimitriosa et al. 2013). Moreover, the scientific debate on performance measurement systems (PMS) introduced in public and non-profit organizations is quite active (Micheli and Kennerly 2005; Trkman 2010; Bititci et al. 2012; Pekkanen and Niemi 2013; Bourne et al. 2014). So, the research community on simulation is stimulated to develop prediction tools to be used in conjunction with pre-existing process monitoring tools.

Non-profit and public organizations generate most of their income from public funds and have to account for several, sometimes conflicting, stakeholders. Their public mission is affected by budget cuts that are becoming more and more severe in several countries. Even educational organizations are becoming business-oriented in their core business despite the social-cultural inspiration at their basis. So, evaluating process-related performance measures in this sector (Jääskeläinen et al. 2015) is first viewed as a means for achieving increased targets of efficiency in the operational behavior of educational institutions as well as steering the allocation of scarce resources. Second, performance evaluation also allows pursuing purposes of accountability and transparency as required by recent laws. To align the business process of a public educational institution, located in Southern Italy, with enterprise like performances a PMS has been formally adopted. As well pointed out in (Han et al. 2009), whenever target organizational processes that need improvement are identified through a macro process analysis, then performance prediction by a micro process analysis using simulation may be worthy. Specifically, with reference to the above PMS we investigate by simulation one of the key (operational) processes in our University Department: the purchasing process. Our aim is to predict if specific objectives assigned to the Department by the Central Administration can be achieved and/or to what extent introducing different organizational set-ups may be necessary to accomplish the above targets.

The paper is organized as follows. In the problem statement, we first provide an overview of recent laws

that have stimulated the introduction of PMSs in Italy and then describe the context of our case study. The Simulation section focuses on input and output modeling issues. The actual case study is presented in the subsequent section with numerical results on some what-if analysis. Conclusions are drawn at the end of the paper.

PROBLEM STATEMENT

With respect to the current public management trend known in literature as New Public Management (Brignall and Modell 2000; Pollitt and Bouckaert 2011), in Italy some recent laws have represented an important “formal” acceleration towards decision supporting in the public sector through the implementation of performance measurement systems.

The common requirement of the Italian normative consists in implementing a strategic performance system to measure, evaluate and improve system performance in terms of efficiency, effectiveness, quality, outcome and customer satisfaction. Similar to the process management logic applied in profit-oriented private companies, the public management reform seeks to apply managerial criteria respecting the general non-profit finality of the public sector.

More specifically, the Italian legislative decree n°150/2009 imposes the implementation of a so-called “performance cycle” in all public organizations. Special government bodies such as the CiVIT (*Commissione per la valutazione, la trasparenza e l'integrità delle amministrazioni pubbliche* - an independent Commission for the evaluation, transparency and integrity of Italian public administrations and which today is known as ANAC, *Autorità Nazionale Anticorruzione* - the national anti-corruption Authority) and ANVUR (*Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca* - an Agency for the evaluation of the activities of Italian public universities and other research bodies) work towards this specific goal and, among the other things, foster:

- transparency and integrity to prevent corruption, with a specific Section dedicated to transparency and integrity;
- improvement of performance management;
- quality of services.

In particular, in 2015 ANVUR issued the guidelines on the “integrated management of the performance cycle in Italian public universities” (ANVUR 2015). According to these guidelines, every university must adopt a system to:

- measure and evaluate the performance of their organization as a whole, as well as the individual performance pertaining to administrative and technical employees;
- establish the method, timelines, processes, instruments and involved subjects.

What was once an opportunity is now a formal requirement in all Italian public universities that may boost or reduce the amount of public funding.

Context of the Study

This study refers to the University of Calabria (www.unical.it) which is located in Southern Italy and counts more than 30,000 students. The document describing the Performance Measurement and Evaluation System of the University of Calabria was approved by its Board of Directors in July 2015. According to this system, whose implementation should be completed by the end of 2016, the so-called “performance cycle” consists in a set of activities aimed at guaranteeing the direction, coordination, control and reporting of university activities. It is composed by the following five phases:

1. medium-long term planning and strategy definition;
2. short-term objectives programming and indicator definition (i.e. specific, measurable, achievable, relevant, and trackable);
3. performance measurement and analysis;
4. organizational and individual performance evaluation and analysis;
5. reporting and transparency.

Within phases 3 and 4, which are currently under implementation, the Board of Directors has assigned a set of first-level objectives pertaining to the administrative management of its individual (and autonomous) Departments. The measurement indicators provided are the following:

- incidence of delay in invoice payment;
- reduction of the average travel refund;
- timely revenue regularization;
- cash balance;
- increase of foreign funding share;
- increase of auto-funding ratio on government financial funding;
- increase of auto-funding trend;
- improve of revenue trend;
- amount of revenue;
- percentage of transfer revenue over total revenue;
- percentage of internal transfer revenue over total revenue;
- full cost reduction of processes developed in different areas.

As a result of this step, every single Department is bound to define and assign second-level (operational) objectives to its employees and then measure their performance. To support the Chair of the Department in this task, a simulation model may be used as a twofold *in vitro* lab. On one hand, it can support the evaluation of employee performance under the current or future organization (and assign bonuses eventually) w.r.t. the assigned objectives. On the other, should the available resources fail to reach a pre-defined target level, it may be used to adjust the objectives and targets to the meet the potential of the Department’s actual human resources. This is a crucial point in the production of public services, since in Italy the acquisition of human resources is currently restricted by the law.

Here we focus on the modeling and simulation of the purchasing process at the Department of Informatics, Modeling, Electronics and System Engineering

(DIMES) of the University of Calabria. Today's national and internal rules make purchasing activities reasonably standardizable. That is the reason of our choice, along with the fact that reducing the payment time of invoices related to the purchasing process is an objective assigned to the DIMES for the year 2016.

The Purchasing Process

In order to favor comprehension of the Department's purchasing process, the overall logic is illustrated by the flowchart in Figure 1. A step-by-step description follows for each block in the flowchart with respect to the activities to be performed, the (human) resources involved in doing so and decisions to be made.

A purchase request is generated by a faculty member to notify the Department's Purchasing Office of items he/she needs to order, along with the quantity and the research funds to be used for such purpose. The request is actually prepared with the support of personnel from the Purchasing Office and not only dispatched to this office after it has been filled in by the faculty member. Consequently, this stage also accounts for the time during which interaction between the two parties takes place if the request contains missing and/or unclear details. Once the request is complete, staff from the Purchasing Office first verifies if the goods/services are available on the Italian Public Administration Marketplace, also known as MEPA. On this digital marketplace (e-procurement) public administrations purchase goods/services, as long as their cost is below a prefixed European threshold. Goods/services are chosen among those offered by suppliers that have been vetted and authorized to post their catalogues on the system. If the goods/services are available on MEPA, then one of the two situations may occur:

- the cost is greater than €4,000 (VAT excluded);
- the cost is less than or equal to €4,000 (VAT excluded).

In the former case, a request for proposal (RFP) is generated on the MEPA portal. This document triggers a sealed-bid procurement procedure through which the Purchasing Office informs the potential MEPA suppliers of the description, technical details, terms and conditions of the goods/services to be procured. Suppliers must submit their offer before the proper time interval (usually, 10 days) has expired. Bid opening, examining and evaluation are carried out by a commission appointed by the Chair of the Department. In particular, the commission usually provides a score for the technical aspects, while MEPA assigns a score for the economic aspects of the bids. If the contact is awarded, than the related legal document is drawn up; otherwise, the procedure overrides any other step and the process is terminated without a winner. In the latter case, the goods/service are chosen from the catalogue of one of the MEPA suppliers and a purchase order (PO) is issued by the Purchasing Office immediately after.

When the goods/services are not available on MEPA, one of the following situations may occur with respect to the extra-MEPA options:

- there is a single supplier;
- there are multiple suppliers.

If there is a single supplier, then only one quotation is requested and a PO is issued by the Purchasing Office. If there are multiple suppliers, but the cost of the goods/services is less than or equal to €4,000 (VAT excluded), then, again, only one quotation is necessary and a purchase order is issued by the Purchasing Office. If the cost is greater than €4,000 (VAT excluded), at least three quotations are acquired and the "best" among these offers is selected. Whatever be the cost of the goods/services of interest, the Purchasing Office completes all extra-MEPA purchases by issuing a PO.

At this point, the order is placed and the lead time between goods/services delivery may vary from time to time. According to the current practice, the Purchasing Office contacts on a, more or less, regular basis the supplier to receive updates on the scheduled due date. If the goods/services have not been delivered, the office urges the supplier to act quickly in order to meet the delivery date or minimize the delay time when already overdue. Once delivered, the Purchasing Office checks to see if the supply is compliant with the order, otherwise the supply is returned to the supplier for replacement/repair. In this stage, the correctness of the related invoice is verified as well and followed by a correction request eventually. After the final approval by the faculty member-funds holder, other formal details are verified in the last stage by the Purchasing Office. The payment order is then prepared, printed, controlled and then signed by those delegated with procurement authority and transmitted to the bank. The purchasing process is terminated, unless the purchased items require being added to the inventory.

SIMULATION

It is easy to recognize from the flowchart in Figure 1 that the state dynamics of the process under examination are determined by the occurrence of some well identified fundamental events, such as request, order, delivery and payment. As a response to the triggering effect of these events, specific activities are performed. Therefore, events and activities determine the evolution over time of the performed payments, by which one should measure the throughput of the entire process. Due to the unavoidable presence of randomness in event occurrence and activity duration, discrete-event (stochastic) simulation (Law and Kelton 2000) appears to be the most appropriate methodology for a quantitative analysis aimed at predicting the performance of the organization policies for the process of interest. If the model had to account for cooperation among staff units or other forms of interaction-based working methods, an agent-based stochastic simulation (Wilensky and Rand 2015) would have been the most natural choice.

Several commercial tools (e.g. Arena 2006; Process 2015) support the implementation of the above flowchart model, as well as its time behavior reproduction.

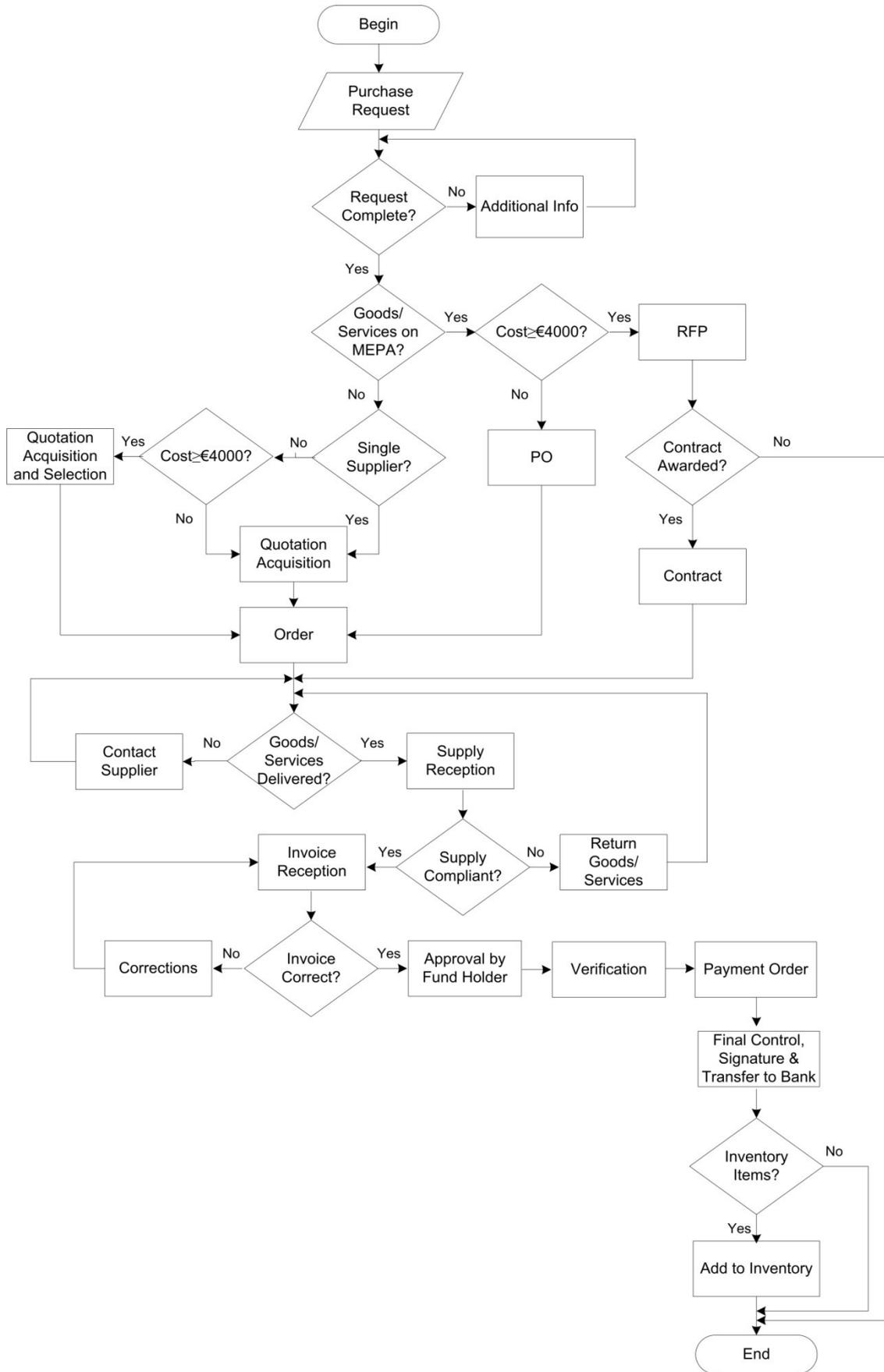


Figure 1: The Purchasing Process

Despite this, simulation input and output analysis is still challenging when facing processes where limited resources play a significant role in determining the process throughput. Hence, among all the steps carried out within our sound and thorough simulation study, here we focus on the statistical analysis of both the simulation input and output data.

In particular, let $\lambda(t)$ be the average number of “purchase requests” arriving in the time unit t (i.e. one week in our case). We partition the real-event data into a suitable sequence of weekly subintervals over a one-year horizon to model the peaks (i.e. before/after breaks and holiday seasons) and troughs (i.e. during breaks and holiday seasons) that can significantly impact on system performance. We then adopt point and interval estimators for the cumulative intensity (or mean value) function (1) of the purchase request events that occur over time in a non-stationary process evaluated at the weekly endpoint t_2 , as suggested by (Leemis 1991).

$$\Lambda(t_1, t_2) = \int_{t_1}^{t_2} \lambda(t) dt \quad (1)$$

Basically, we assume that a non-homogeneous Poisson process (NHPP) with a piecewise constant intensity (rate) function is appropriate enough to model the series of events that occur over the weeks in a non-stationary fashion (see Table 1). Algorithms presented by (Leemis 2004) are then used in the simulation experiments to generate purchase order arrival times from the estimated NHPP.

Table 1: The Piecewise Constant Intensity Function for Modeling the Arrival of Purchase Requests

	Rate [arrivals/week]	Duration [week interval]
λ_1	4	1-8
λ_2	13	9-30
λ_3	8	31-52

To complete the input modeling analysis, other types of distribution probability functions have been identified (with Arena’s Input Analyzer) for the second major sources of uncertainty, i.e. the activity durations:

- PO processing times are well captured by a 2-order Erlang distribution with mean value equal to 19.1 and shifted to the right by 0.999 ($0.999 + Erlang(19.1,2)$);
- RFP preprocessing and processing times profiles are well fitted by a 2-order Erlang distribution with mean value equal to 8.22 and shifted to the right by 8.5 and a Beta-based distribution with shape 1 equal to 0.857, shape 2 equal to 1.65 and shifted to the right by 10, respectively ($8.5 + Erlang(8.22,2)$ and $10 + 142 * Beta(0.857,1.65)$);
- Extra MEPA processing time profiles agree with a 2-order Erlang distribution with mean value equal

to 25.6 and shifted to the right by 5 ($5 + Erlang(25.6,2)$);

- Final control and transfer of payment documents to the bank are modeled by means of a Normal density function with mean value equal to 4.54 and standard deviation equal to 2.28 ($Normal(4.54, 2.28)$).

As for simulation output analysis, we used a first set of results to perform validation, i.e. assess if the real purchasing process is accurately represented by the simulation model. Besides discussing the structural assumptions and data assumptions with the key figures of the Department’s Purchasing and Administration Offices, model input-output transformations have been compared to the corresponding input-output transformations for the real system.

As a result of both of these validation activities, one can appreciate the capability of the simulator to mirror the real system performance by observing the 95% interval estimates vs the real figures in Table 2 and the simulated vs the real trend in Figure 2. In the former case, although the X-MEPAs real measure (i.e. 147) is right-adjacent to the interval returned via simulation, it is still a good result if one considers the overall degree of uncertainty intrinsic in the specific X-MEPA process due to the difficulty in measuring the time effort required by the Commission. As for the latter case, the real (blue) trend in Figure 2 refers to the unique real throughput trajectory available, whereas the simulated (red) trend is an estimate of the average throughput behavior. The degree of matching between the two above trends is satisfactory enough, especially after the first weeks (once transient behavior has died out).

Table 2: Real vs Simulated N° of Purchase Requests

Source	Purchase Requests		
	RFPs	POs	X-MEPAs
2014 Records	35	277	147
Simulator	[31-38]	[265-280]	[134-146]

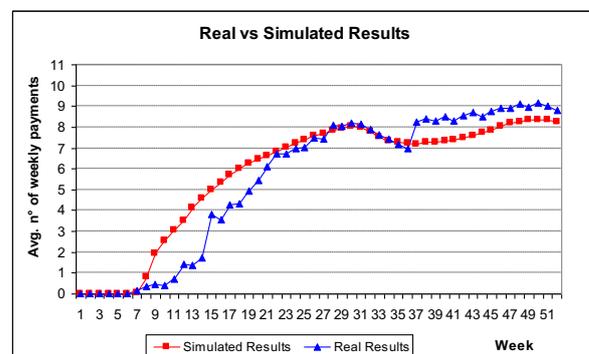


Figure 2: Real vs Simulated Trend of the Average N° of Purchases Completed (Throughput) per Week

CASE STUDY

The purpose of this case study is to compare the “as is” organization of the Department’s purchasing process with its “to be” organization based on policies compliant with national compulsory anticorruption and transparency service regulations. In particular, within the wider objective of rationalizing the public expenditure and reorganizing the administrative structure, here we intend to perform two what-if experiments. In the first, the idea is to lower the €4,000 disclosure threshold referred to in the description of the purchasing process. As a result, more RFPs, rather than POs should be generated on the MEPA portal, thus, preventing the use of working policies in the public sector that lack transparency and encourage, in some sense, favoritism. In the second, in shifting from one set-up to another, we wish to account for non-secondary organizational issues in the purchasing process such as introducing a new policy when carrying out final controls and transferring payment orders to the bank.

The simulation experiments for both cases have been carried out under Rockwell’s Arena simulation package (version 11) and run on a personal computer equipped with an Intel® Core™2 Duo 1.58 Ghz processor and 2.93 gigabytes of RAM. The models in Arena include VBA (Visual Basic® for Applications) blocks that allow inserting user-defined code. In our study, these blocks are used to interact with worksheets under Microsoft® Excel in order to record the output data produced by the simulator and generate 95% interval estimates. All the experiments share the same computational effort (i.e. 30 runs for each 1-year scenario), the same number of resources (i.e. 1 unit in the Administration Office, 1 unit in the Purchasing Office and 1 unit in the Payment Office) and the same *modus operandi*.

The “Transparency” What-if

In this first what-if analysis, we consider the importance of the so-called disclosure threshold of the goods/services available on MEPA. As previously stated, the value of this threshold determines whether the goods/services of interest should be purchased via a sealed-bid RFP or chosen directly with a PO from the catalogue of one of the MEPA suppliers. This value is currently set at €4,000. However, in order to increase transparency, we believe worthy investigating the effect of lowering the above disclosure threshold. Since in our Department only 7.6% of the purchases are carried out according to the RFP option (see column 2 in Table 3), the point becomes whether or not the organization and personnel can cope with the greater operational effort required by a decrease in the above threshold.

Let us assume that one of the Department’s operational objectives consists in decreasing the threshold from €4,000 to another value between €3,000 and €1,000. As a result, the RFP-based purchases will

go from 7.6% to some value between 10.7% and 22%, respectively (see columns 3 to 5 in Table 3).

Table 3: Composition of Purchase Types according to Disclosure Threshold

Type	Disclosure Threshold			
	€4,000	€3,000	€2,000	€1,000
X-MEPAs	32.0%	32.0%	32.0%	32.0%
POs	60.4%	57.3%	55.3%	46.0%
RFPs	7.6%	10.7%	12.7%	22.0%

As one may see from the simulation results reported in Table 4, fixing the disclosure threshold to either €3,000 or €2,000 is well-supported by the overall purchasing process: only small changes occur in the average number of purchases completed per week. On the other hand, if the threshold value drops to €1,000 (or below), then the overall performance of the process will drop considerably as well: the average value of the number of purchases completed per week will go from 8.37 to 6.54. This is probably due to the fact that the moderate-high level of utilization of one of the key resources in the purchasing process (i.e. an average 75% for the unit in the Purchasing Office) inevitably drives this resource to become the system bottleneck. Thus, the level of human resource utilization cannot be disregarded when tuning the value of the disclosure threshold.

Table 4: Average N° of Purchases Completed per Week for a Range of Disclosure Thresholds

Disclosure Threshold			
€4,000	€3,000	€2,000	€1,000
[8.28-8.45]	[7.96-8.20]	[7.56-8.10]	[6.32-6.75]

The “Final Control and Delivery Policy” What-if

In this second what-if analysis, we consider the effect of introducing a new policy when carrying out the final controls and transferring the Department payment orders to the bank. As of today, control and transfer occurrences depend on a variety of contingencies (e.g. deadlines, priorities and personnel availability), rather than a fixed scheduling policy.

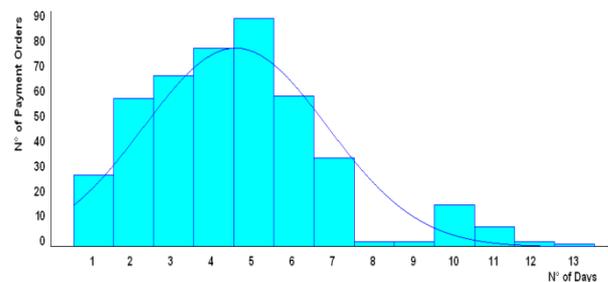


Figure 3: Duration of the Time to Control and Transfer Payment Orders to the Bank

In our actual case, once payment orders are authorized, then signing, control and transfer operations are carried out according to the profile illustrated in Figure 3, which bears an average value of 4.54 days, a 2.29 standard deviation and a 0.89 skewness. The profile has been obtained with Arena's Input Analyzer and fitted to a Normal shape (with skewness set to zero).

Time-based or batch-based policies are two of the new possible scheduling options. In the former case, a payment order is controlled and transferred only if it arrives within a given time interval. In the latter, payment orders are collected, controlled and transferred to the bank only when the number of orders in a batch reaches a target maximum.

Table 5: Duration of the Overall Completion Time of the Purchasing Process

Lead Time (days)			
Policy	X-MEPAs	Pos	RFPs
current	[27.6-34.1]	[25.1-30.9]	[53.3-59.0]
time-based	[26.1-34.4]	[23.4-31.5]	[51.5-59.7]

Here we focus on a 24-hour time-based policy according to which payment orders are controlled and transferred (by the unit working in the Administration Office) if they arrive duly approved before 1:00 p.m. of every day; otherwise, they are controlled and transferred the day after.

As one may expect, the Department's purchasing process benefits from introducing a time-based scheduling policy: the average control and transfer time decreases from 4.54 days (1 day = 7 working hours) to 3.22 hours and the resulting shape in Figure 4 is only slightly skewed to the left (-0.13).

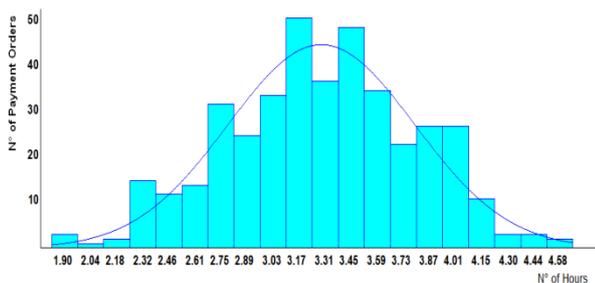


Figure 4: Duration of the Time to Control and Transfer Payment Orders to the Bank with New Policy

The great reduction from approximately 30 to 3 working hours is certainly due to the daily-based control and transfer mechanism of the new policy against the contingency-based decision pertaining to the previous practice. As a matter of fact, the central (average) value of the interval estimates for the lead time of the overall purchasing process reported in Table 5 is shifted towards the left (i.e. reduced), although the whole interval has become slightly larger.

CONCLUSIONS

It has been shown that computer simulation of the "purchasing process" in a University Department is both appropriate and effective for predicting process performance and the benefit of innovation. Validation has been successfully carried out to assess the reliability of the chosen data modeling method for input events (purchase requests). Then a what-if analysis has been presented to illustrate the practical use of the simulator when pursuing increased levels of efficiency in terms of both order cycle times and process transparency. In particular, the latter could be achieved through a sustainable reduction of the threshold level enabling a specific procedural variant within the purchasing procedure. The impact of human resources availability and utilization on the performance of the whole process could also be predicted to drive a rational resource allocation over the set of operational activities. More generally, this study should encourage the adoption of discrete-event simulation as the most appropriate prediction tool aimed at supporting the integration of strategic plans and performance targets with operational processes in the public sector.

REFERENCES

- ANVUR. 2015. "Linee Guida per la Gestione Integrata del Ciclo della Performance delle Università Statali Italiane". <http://www.anvur.org/attachments/article/833/Linee%20Guida%20Atenei.pdf> [Accessed January 12, 2016].
- Arena – Version 11.00.00-CPR 7 Copyright © 2006 Rockwell Automation Technologies, Inc.
- Bititci, U.; P. Garengo; V. Dorfler; and S. Nudurupati. 2012. "Performance Measurement Challenges for Tomorrow". *International Journal of Management Reviews* 14, No.3 (Sept), 305-327.
- Bourne, M.; S. Melnyk; U. Bititci; K. Platts; and B. Andersen. 2014. "Emerging issues in performance measurement". *Management Accounting Research* 25, No.2 (Jun), 117-118.
- Brignall, S. and S. Modell. 2000. "An Institutional Perspective on Performance Measurement and Management in the 'New Public Sector'". *Management Accounting Research* 11, No.3 (Sept), 281-306.
- de Boer, L.; M. Ebben; and C. Pop Sitar. 2003. "Studying Purchasing Specialization in Organizations: a Multi-agent Simulation Approach". *Journal of Purchasing & Supply Management* 9, No.5-6 (Sept-Nov), 199-206.
- Dimitriosa N.K.; D.P.Sakasa; and D.S.Vlachos. 2013. "Analysis of Strategic Leadership Simulation Models in Non-profit Organizations". *Procedia - Social and Behavioral Sciences* 73, (Feb), 276-284.
- Greasley, A. 2006. "Using Process Mapping and Business Process Simulation to Support a Process-Based Approach to Change in a Public Sector Organization". *Technovation* 26, No.1 (Jan), 95-103.
- Gregoriades, A. and A. Sutcliffe. 2008. "A Socio-Technical Approach to Business Process Simulation". *Decision Support Systems* 45, No.4 (Nov), 1017-1030.
- Han, K.W.; J.G. Kang; and M. Song. 2009. "Two-stage Process Analysis using the Process-Based Performance Measurement Framework and Business Process

- Simulation". *Expert Systems with Applications* 36, No.3 (Apr), 7080-7086.
- Haysa, M.A. and M. Bebbington. 2000. "Simulation in Public Sector Management: a Case Study". *International Transactions in Operational Research* 7, No.4-5 (Sept), 465-486.
- Hlupic, V. and G.J. de Vreede. 2005. "Business Process Modeling using Discrete-Event Simulation: Current Opportunities and Future Challenges". *International Journal of Simulation & Process Modeling* 1, No.1-2, 72-81.
- Jääskeläinen, A.; A. Lönnqvist; and H.I. Kulmala. 2015. "Designing a Performance Measurement System to Support Outsourcing Decisions in a Finnish University". *International Journal of Public Sector Performance Management* 2, No.3, 237-252.
- Law, A.M. and W.D. Kelton. 2000. *Simulation Modeling and Analysis* 3rd edn., McGraw-Hill, New York, NY.
- Leemis, L.M.1991. "Nonparametric Estimation of the Intensity Function for a Nonhomogeneous Poisson Process". *Management Science* 37, No.7 (July), 886-900.
- Leemis, L.M. 2004. "Nonparametric Estimation and Variate Generation for a Nonhomogeneous Poisson Process from Event Count Data". *IIE Transactions* 36, No.12 (Dec), 1155-1160.
- Melao, N. and M. Pidd. 2006. "Using Component Technology to Develop a Simulation Library for Business Process Modeling". *European Journal of Operational Research* 176, No.1 (Jul), 163-178.
- Micheli, P. and M. Kennerly. 2005. "Performance Measurement Frameworks in Public and Non-profit Sectors". *Production Planning & Control* 16, No.2 (Feb), 125-134.
- Pekkanen, P. and P. Niemi. 2013. "Process Performance Improvement in Justice Organizations - Pitfalls of Performance Measurement". *International J. Production Economics* 143, No.2 (Jun), 605-611.
- Pollitt, C. and G. Bouckaert. 2011. *Public Management Reform: a Comparative Analysis* 3rd edn., Oxford University Press, Oxford.
- Process 2015 – Version 15.2.2.1608 iGrafx, LLC.
- Robinson, S.; C. Worthington; N. Burgess; and Z.J. Radnor. 2014. "Facilitated Modelling with Discrete-Event Simulation: Reality or Myth?". *European Journal of Operational Research* 234, No.1 (Apr), 231-240.
- Saldívar, J.; C. Vairetti; C. Rodríguez; F. Daniel; F. Casati; and R. Alarcón. 2016. "Analysis and Improvement of Business Process Models using Spreadsheets". *Information Systems* 57, (Apr), 1-19.
- Shafer, S.M and T.L. Smunt. 2004. "Empirical Simulation Studies in Operations Management: Context, Trends, and Research Opportunities". *Journal of Operations Management* 22, No.4 (Aug), 345-354.
- Trkman, P. 2010. "The Critical Success Factors of Business Process Management". *International Journal of Information Management* 30, No.2 (Apr), 125-134.
- Van der Aalst, W.M.P. 2010. "Business Process Simulation Revisited". In *Enterprise and Organizational Modeling and Simulation*, J. Barjis (ed.). *Lecture Notes in Business Information Processing* 63, Springer-Verlag, Berlin, 1-14.
- Wilensky, U. and W. Rand. 2015. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems With Netlogo*, MIT Press, Cambridge, MA.

AUTHOR BIOGRAPHIES



PASQUALE LEGATO is an Associate Professor of Operations Research in the Department of Informatics, Modeling, Electronics and System Engineering (DIMES) at the University of Calabria, Rende (CS, Italy). He has been a member of the Executive Board of the University of Calabria as well as university delegate for the supervision of associations and spin-offs from the University of Calabria. He has been involved in several EEC funded research projects aimed at the technological transfer of simulation based optimization procedures and frameworks in logistics. Currently, he is a member of the INFORMS Simulation Society and is serving as a reviewer for both INFORMS and Elsevier journals. His research activities focus on predictive stochastic models for cyber security, queuing network models, stochastic simulation and the integration of simulation techniques with combinatorial optimization algorithms. His e-mail address is: legato@dimes.unical.it and his web-page can be found at <http://www.info.dimes.unical.it/legato>.



LIDIA MALIZIA is the Administrative Manager of the Department of Informatics, Modeling, Electronics and System Engineering (DIMES) at the University of Calabria, Rende (CS, Italy). She graduated in Business Administration and received a PhD in Management and Economics of Public Administrations from the above university. She is also a Certified Public Accountant and Auditor. She has a fifteen-year working experience on management and accounting in public administrations. Her current research interests include emergent management and accounting models in Italian public universities. Her e-mail address is lmalizia@dimes.unical.it.



RINA MARY MAZZA is the Research Manager of the Department of Informatics, Modeling, Electronics and System Engineering (DIMES) at the University of Calabria, Rende (CS, Italy). She graduated in Management Engineering and received a PhD in Operations Research from the above university. She has a seven-year working experience on knowledge management and quality assurance in research centers. She has also been a consultant for operations modeling and simulation in container terminals. Her current research interests include discrete-event simulation and optimum-seeking by simulation in complex systems. Her e-mail address is: rmazza@dimes.unical.it.

Path dependence in hierarchical organizations: The influence of environmental dynamics

Arne Petermann
Institute for Quality and Management, BAGSS
Konrad-Zuse-Str 3a, 66115 Saarbrücken, Germany
E-mail: arne.petermann@iqm-bagss.de

Alexander Simon
Berlin University of Professional Studies
Katharinenstraße 17-18, 10711 Berlin, Germany
E-mail: alexandersimon90@googlemail.com

KEYWORDS

agent based modelling, social simulation, path dependence, path breaking, organizational studies, business strategy, technology strategy

ABSTRACT

The following paper will describe how path dependent hierarchical organizations are affected by a changing environment. The results of current 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 is changing, the existing model will be enhanced. The objective is to simulate the impact of external influences to the emergence of norms within an organization.

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. If an organization fails to do so, it may fall back or even be eliminated from the competition. This comes with a high risk, especially 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 technological change on the TV market from the CRT displays to the new LCD-based flat screen technologies. In fact, Loewe still builds CRT displays. The result is an imbalance of supply and demand, because most customers are not interested in those TV's any more. 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 the investor Stargate Capital bought Loewe and made some serious changes. But their previous ignorance almost led them into bankruptcy.

Nokia on the other hand 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 and Samsung captured the market applying the latest technologies. By now the mobile phone division of Nokia has been bought by Microsoft. 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) which for reasons of simplicity was built on the assumption of a stable environment will be extended with a new variable one or the other 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 dug 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. In that time this new technology was clearly a superior solution than the incumbent. These keyboards, however, were not able to become a serious competitor to "QWERTY"-keyboards. David examined in detail the self-reinforcing mechanisms that led to the domination of the established keyboards by the inferior QWERTY solution.

Based on David's findings, Arthur (1989) used a poly urn model to analyze the self reinforcing mechanisms discovered by David in a more formal way. In his model two technologies (A and B) are entering the market at the same time and compete for the adoption by customers called agents. At the beginning both technologies have the same chances 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.

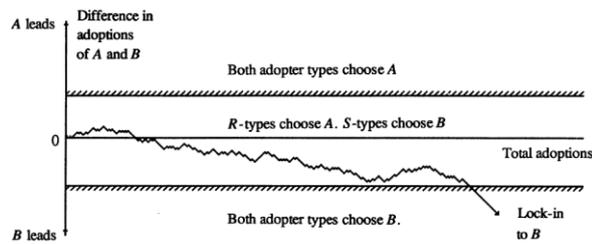


Figure 1: Increasing returns adoption: a random walk with absorbing barriers

David and Arthur stress that a technology can become dominant even when it is inferior in terms of its long term value to the system.

Transform path dependence to an 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 developed a model which describes this advanced concept of path dependence. In this model the path is split into three different phases. Figure 2 (Sydow et al., 2009: 692) shows the concept of this model.

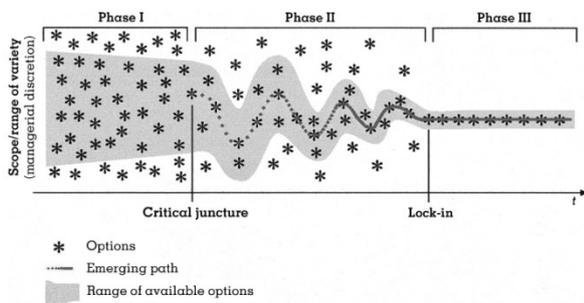


Figure 2: The Constitution of an Organizational Path

Phase 1: Preformation Phase

In this phase the decisions the participants are able to make are relatively open. Influences at this point can 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". 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". The self-reinforcing effects that are triggered by these small events narrow the path and limit future choices within the organization.

Phase3: Lock-In Phase

The reinforcing effects have now taken the lead and reduced the scope of choices drastically. 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, 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 organizations that undergo path depended processes. In historical analysis scholars have shown many examples of organizations that were able to adapt in the light of changing environment, while others are stuck in a lock-in state, unable to change, even when the necessity to change became obvious from an outside perspective. Our model aims at describing hierarchical organizations, that undergo a path depended 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 (Harrison 2007: 1232). When complexity and non-linearity of social systems make it hard or impossible to develop mathematical equations, 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 insights and first assumptions for other social research methods.

The basic model

The basics of this research is the simulation model M1 Petermann et al. (2012) developed in their 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, which 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 cases (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 his 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

Enhancing this model further, we now implement an external impact into 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.

$$FTA_M(x,y,z) = E_1 * f_{M,c}(x) + i(y) * li + s(z) * ei \quad (4)$$

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 other variables in the equation are generated by the simulated organization itself, ei is triggered from an external source. When there is no external impact, ei is equal to 1 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. Is 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.

SIMULATION

To simulate the described external impact, we need to specify in what way and when the variable ei should change. The first condition, we need to break the path and the path must be locked-in. That means that a dominant norm exists in the simulation model. The lock-in state in model M1 is defined by Petermann et al. (2012: 195) as minimum 500 of ticks with a spread > 0.9 if B is dominant (spread < -0.9 if A is dominant). Furthermore, a definition for breaking a path is also needed. The model M1 defines no values for that, so we assume a path is broken when spread < 0.5 , when B was the current norm in the company and a spread > -0.5 when A was the current norm. This means that less than 75% of all company members adopted a norm. The last variable is the leadership impact. This is set to 1, to have an impact from that side. The defined values for lock-in and leadership impact are assumption and not empirical researched values.

The variation of the external impact is possible in two ways. Either the intensity can be varied or the amount of time (number of ticks) the impact is present in the system. To get usable data from the simulation model, only datasets with a lock-in at B before the external impact is triggered are used for analysis purposes. Otherwise it is not possible to see a behavior for one norm. A simulation for each parameter-set will run 100 times according to the Monte-Carlo method (Law et al. 1991:113).

Run of the 1. Simulation

The change of the external impacts must be further clarified to run the first approach. During the enhancement of the model the following parameters seemed to be valid for a first run. After the first simulation an optimization of the parameters is needed. Maybe a closer look at several parameter areas is interesting.

		intensity (int)				
		1	3	5	7	10
continuity (ticks)	10	10/1	10/3	10/5	10/7	10/10
	40	40/1	40/3	40/5	40/7	40/10
	70	70/1	70/3	70/5	70/7	70/10
	100	100/1	100/3	100/5	100/7	100/10
	130	130/1	130/3	130/5	130/7	130/10

Figure 3: 1. Simulation parameter Setup

The lock-in behavior with leadership impact of 1 and 2 is at 6000 ticks (Petermann et al. 2012:195). This means, that each simulation must run at least 6000 ticks, before the external impact can be triggered. A complete run will last 8500 ticks, and then the system has enough time to reconfigure itself after the external impact. It is not possible to define a number of ticks after the impact, when the system has locked-in again. This basic setup is used for all simulations in this paper; otherwise it is not possible to compare the results properly.

Results of the 1st Simulation

The result of this run is a huge amount of data, which needs to be analyzed. The first intensity parameters from 1 to 3 will not be visualized in this paper, because the maximum possible spread change is from 1 to 0,992 at a random point of time, so with a parameter combination of 130 for continuity and 3 for intensity no connection to the external impact can be identified. The effects that occur at the intensity of 10 are also not visualized they are similar to the graphics that depict the intensity of 7.

Results for Intensity of 1 and 3

No valid differences could be detected, that change the system normal behavior. It is not possible, to force a

path break with all combinations containing the parameter 1 for intensity. The most interesting effect occurs at the parameter intensity between 5 and 7.

Results for Intensity of 5

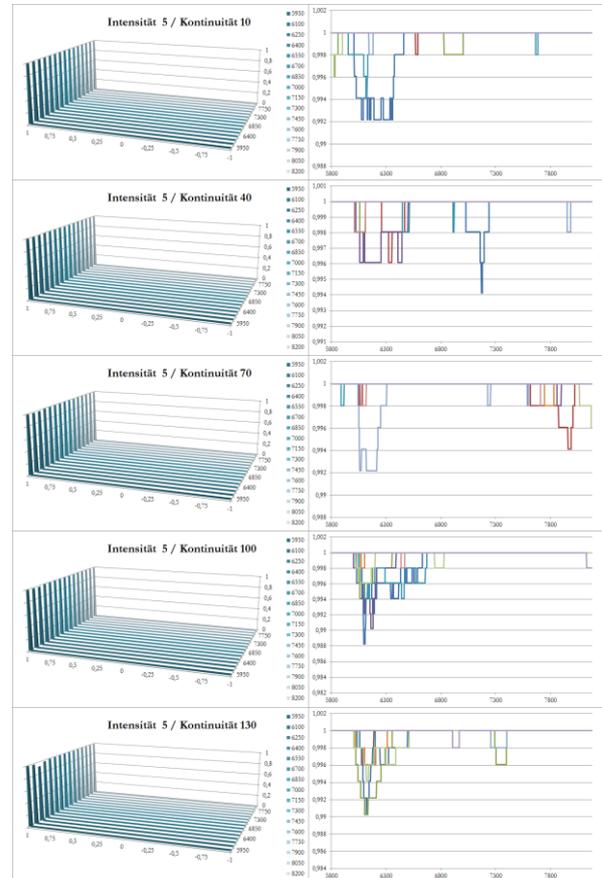


Figure 4: Left: Histogramical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

At this parameter setup the system started to react. With the combination of continuity of 10 until continuity of 70 no valuable reactions are noticeable. However, at a continuity of 100 the system starts to change. The spread is forced to the path breaking direction. Of course, it is only a spread of 0.992, but the events occur exactly at the starting point of the external impact. With this first result it is maybe useful, to increase the continuity over 130 with an intensity of 5 to generate a path break.

Results for Intensity of 7

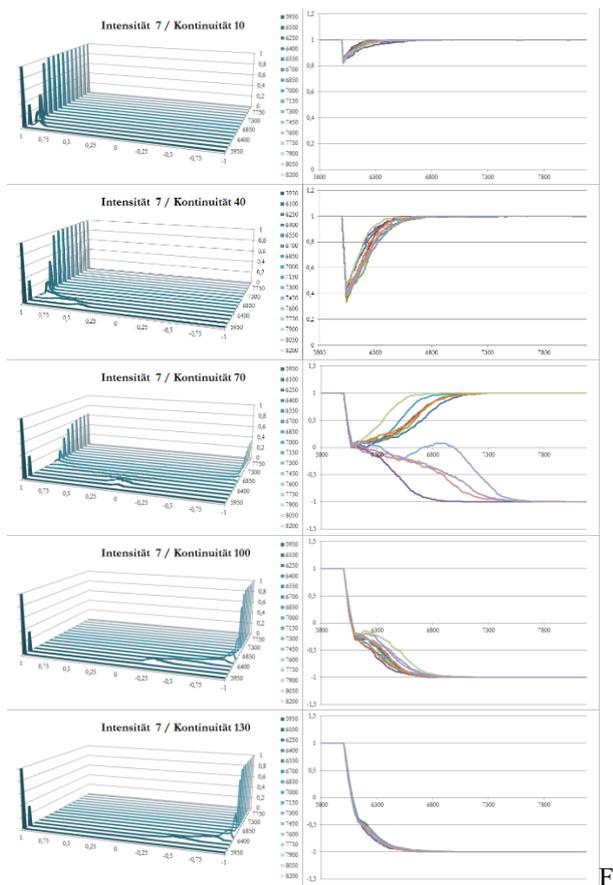


Figure 5: Left: Histogramical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

An intensity of 7 forces the system to break the path for the first time even with low continuities. With a continuity of 10 a significant system behavior is detected, but there is still no path break. This happens for the first time with a continuity of 40 (spread < 0.4). Higher continuities with values of 70, 100 and 130 forced the system to establish new norms.

Results for Intensity of 10

The intensity of 10 behaves nearly like the intensity of 7. With higher amount of continuity path breaks and new path formations are results of the simulation.

Summary of the 1st Simulation

The first simulation run gave first insights to the system behavior of the described model M1. In the following three figures all parameter combinations described in figure 3 are being compared.

The effect at the path is listed in figure 6. As described above, the interesting area is between the intensities of 5

and 7. The probability of a path breaking behavior increases rapidly at this interval. This parameter field will be investigated more closely.

		intensity (int)				
		1	3	5	7	10
continuity (ticks)	10	0%	0%	0%	0%	0%
	40	0%	0%	0%	99%	100%
	70	0%	0%	0%	100%	100%
	100	0%	0%	0%	100%	100%
	130	0%	0%	0%	100%	100%

Figure 6: Pathbreaking probability

Figure 7 describes the probability of new path directions. With a continuity of 70 at an intensity of 7 the probability is 3% higher compared to an intensity of 10. There is, however, a small probability at 100 simulation runs that the result differs from one's expectations. That the system behaves unexpectedly at this point could also be an assumption. A deeper analysis about this could be an interesting question for upcoming research, but it will not find place in this paper.

		intensity (int)				
		1	3	5	7	10
continuity (ticks)	10	0%	0%	0%	0%	0%
	40	0%	0%	0%	0%	0%
	70	0%	0%	0%	35%	32%
	100	0%	0%	0%	100%	100%
	130	0%	0%	0%	100%	100%

Figure 7: new path direction probability

Finally the average spread over all combinations is shown in figure 8. The average was calculated at the last tick of the impact. As expected, the spread changes in the intensity fields of 1 and 3 are out of scope. It's interesting to see that with an intensity of 5 differs with 3%, but that is not according to its continuity.

		intensity (int)				
		1	3	5	7	10
continuity (ticks)	10	1,00	1,00	1,00	0,84	0,82
	40	1,00	1,00	1,00	0,40	0,37
	70	1,00	1,00	0,98	0,05	0,02
	100	1,00	1,00	0,97	-0,23	-0,25
	130	1,00	1,00	1,00	-0,43	-0,45

Figure 8: Average spread, calculated at the last external impact tick

The expected behavior was successfully created in the first simulation run. The path was broken, and new path directions emerged in the system. It is also a validation of the external impact implementation of the model M1. Furthermore interesting results came out of the first run. A second and third simulation run are recommend at this point. In the second simulation a closer look is taken at the system behavior with very high intensities. As described above, the intensities 7 and 10 seem to behave almost equally. To clarify this interpretation, a second simulation was done. The third simulation takes a closer look at the intensity area between 5 and 7. Here the system seems to react very sensitively.

Run of the 2nd Simulation

The finding of the first simulation: “the intensities of 7 and 10 behave almost equally” should be clarified in this simulation. Therefore a simulation with highly overdriven intensities was performed. We assume here that there are no significant different system-behaviors observables, with intensities of 10 or more. To break the path with a continuity of 10 is also part of this simulation. The path was influenced in the first simulation with this continuity, but there was no sustainable effect, like breaking or new formation, at the path. Figure 9 shows the parameter setup for the second simulation.

		intensity (int)		
		13	16	19
continuity (ticks)	10	10/13	10/16	10/19
	40	40/13	40/16	40/19
	70	70/13	70/16	70/19
	100	100/13	100/16	100/19
	130	130/13	130/16	130/19

Figure 9: 2. Simulation parameter Setup

Results of the 2nd Simulation

The figures 10 – 12 show the visualized result of the simulation.

Results for Intensity of 13

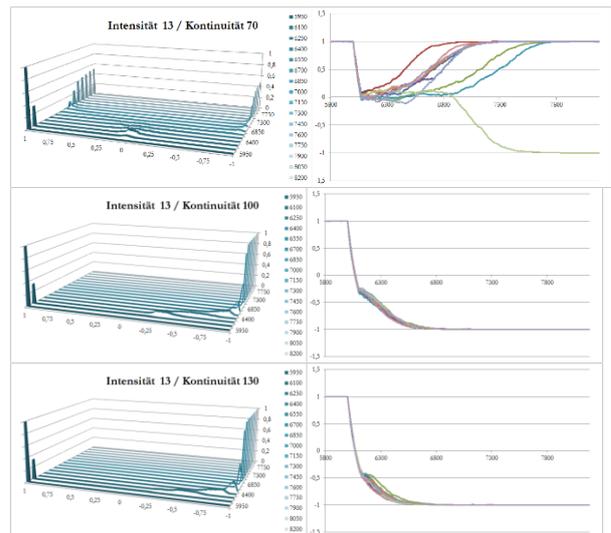
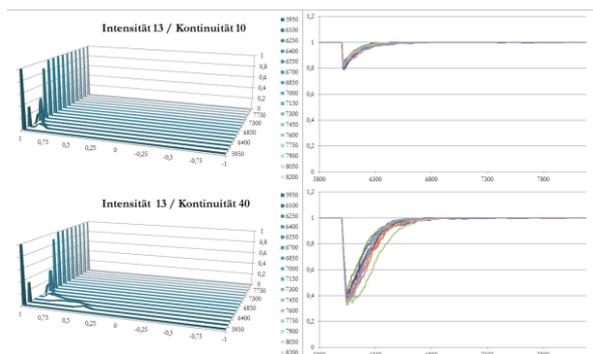


Figure 10: Left: Histogramical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

Results for Intensity of 16

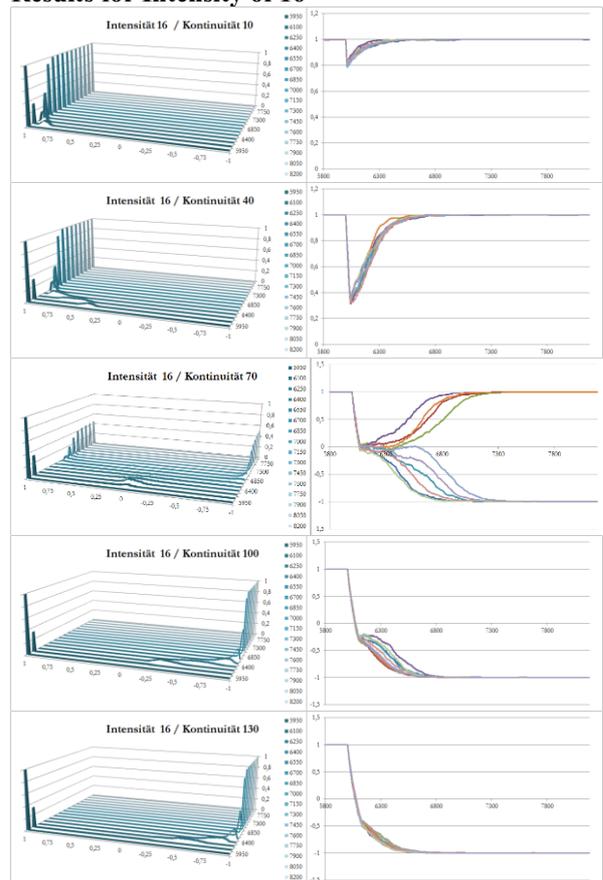


Figure 11: Left: Histogramical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

Results for Intensity of 19

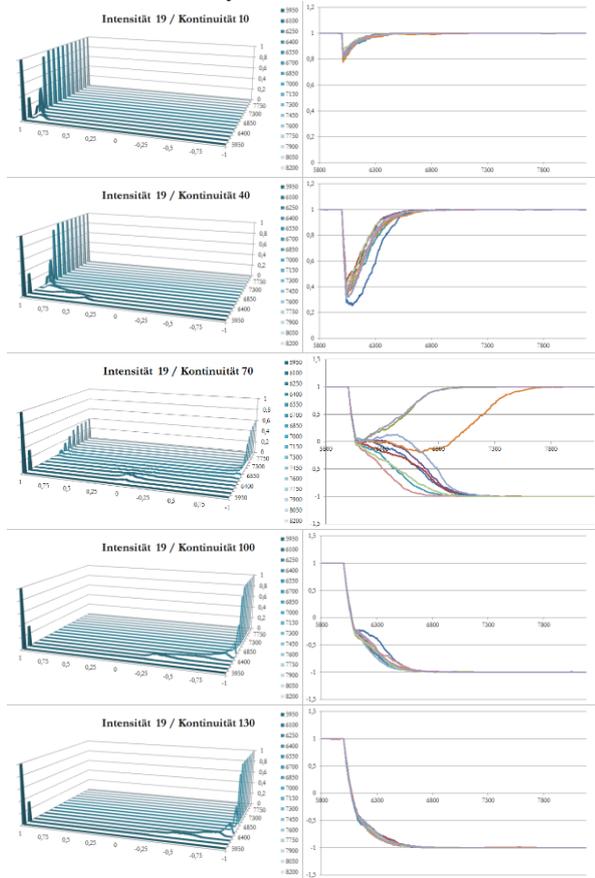


Figure 12: Left: Histogramical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from -1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

The average spreads shall verify the assumption of a constant changing behavior with increasing intensities. The detailed results are shown in figure 13 to 15.

		intensity (int)				
		7	10	13	16	19
continuity (ticks)	10	0%	0%	0%	0%	0%
	40	99%	100%	100%	100%	100%
	70	100%	100%	100%	100%	100%
	100	100%	100%	100%	100%	100%
	130	100%	100%	100%	100%	100%

Figure 13: Pathbreaking probability

		intensity (int)				
		7	10	13	16	19
continuity (ticks)	10	0%	0%	0%	0%	0%
	40	0%	0%	0%	0%	0%
	70	35%	32%	49%	51%	60%
	100	100%	100%	100%	100%	100%
	130	100%	100%	100%	100%	100%

Figure 14: new path direction probability

		intensity (int)				
		7	10	13	16	19
continuity (ticks)	10	0,84	0,82	0,81	0,81	0,81
	40	0,40	0,37	0,35	0,35	0,35
	70	0,05	0,02	0,01	0	0
	100	-0,23	-0,25	-0,25	-0,25	-0,26
	130	-0,43	-0,45	-0,45	-0,45	-0,45

Figure 15: Average spread, calculated at the last external impact tick

Figure 13 shows the pathbreaking probability for the second simulation. All intensities have similar values, except the parameter combination intensity of 7 and continuity of 30 with 99%. The average spread, which is shown in figure 15, only varies from the intensity from 7 to 10. For an intensity of 13 or more the average spread is constant. This result validates the assumption that a continuity of 10 can not trigger a pathbreak in the system, regardless the height of intensity. The most interesting outcomes of this simulation are the values of the new path direction probability (figure 14). The continuities have the same behaviors, except 70. With a higher intensity the new path direction probability increases. Even if the average spread is 0 at a continuity of 70 and intensity of 16 and continuity of 70 and intensity of 19, that means a total balance between the adopted norms A and B is present, the new path direction probability increases about 9% between this two combinations.

Run of the 3rd simulation

This simulation researches the area of the intensities between 5 and 7 as described above. The following table shows the parameter setup for this run.

		intensity (int)		
		5,5	6	6,5
continuity (ticks)	10	10/5.5	10/6	10/6.5
	40	40/5.5	40/6	40/6.5
	70	70/5.5	70/6	70/6.5
	100	100/5.5	100/6	100/6.5
	130	130/5.5	130/6	130/6.5

Figure 16: 3. Simulation parameter setup

Results of the 3rd Simulation

The figures 17 and 18 show the visualized results of the simulation.

Results for Intensity of 5.5

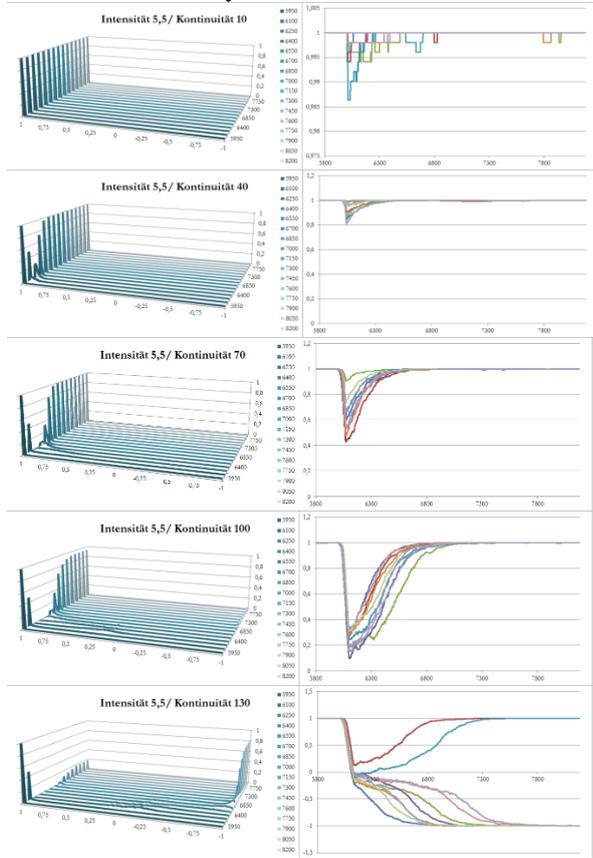


Figure 17: Left: Histogramphical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

Results for Intensity 6

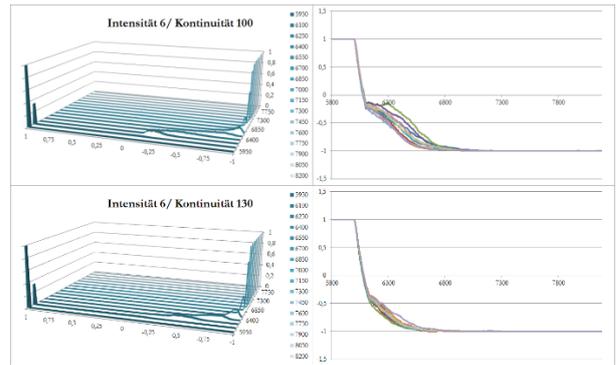
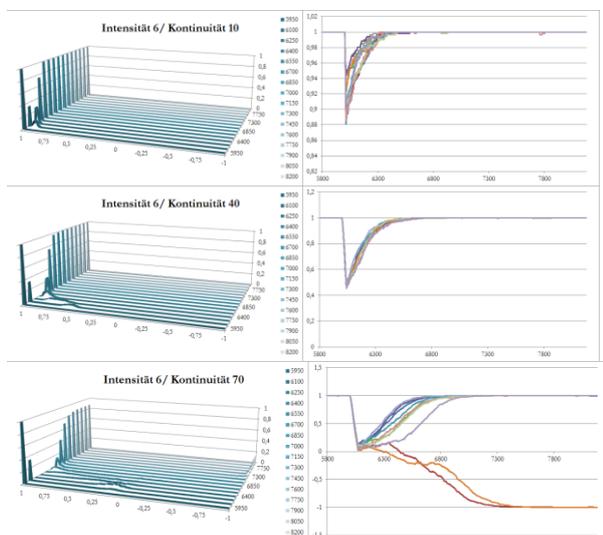


Figure 18: Left: Histogramphical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

Results for Intensity of 6.5

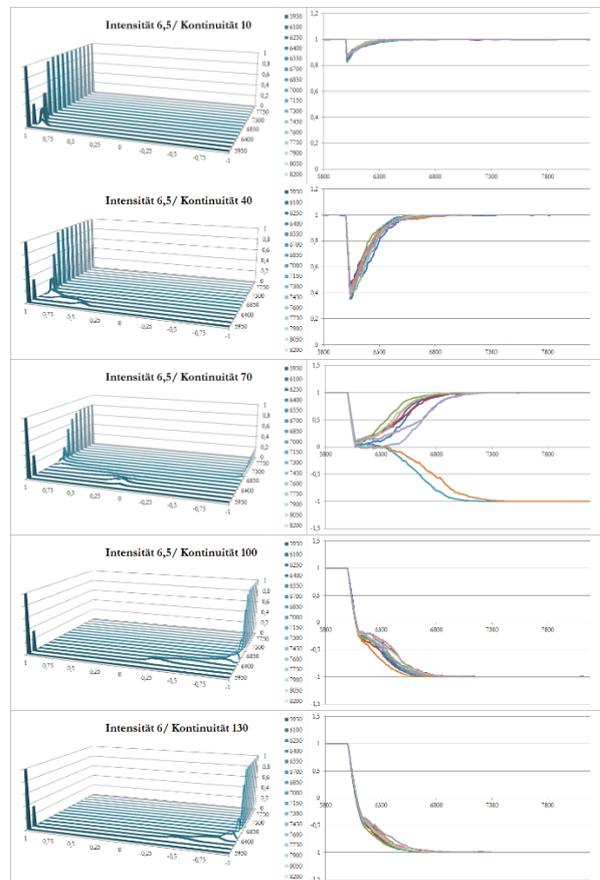


Figure 19: Left: Histogramphical view: probability density, 150 ticks summarized. X axis: spread, Y axis: probability from 0-1, z-Axis: time in ticks, starts counting at 5800 ticks. Right: exemplary first 10 runs. X Axis: ticks, Y-Axis: spread.

In Figure 20 the different pathbreaking probabilities are listed. With an intensity of 6.5 and a continuity of 40 and more the path breaks with a probability of 100%. Compared to the intensity of 7 (figure 6) these two intensities are very close the each other.

		intensity (int)		
		5,5	6	6,5
continuity (ticks)	10	0%	0%	0%
	40	0%	68%	100%
	70	22%	100%	100%
	100	93%	100%	100%
	130	99%	100%	100%

Figure 20: Pathbreaking probability

The new path direction probability increases rapidly from an intensity of 5.5 to 6 with a continuity of 100 from 6% to 98%, shown in figure 21. The highly sensitive area can be bounded between the intensities from 5.5 to 6.

		intensity (int)		
		5,5	6	6,5
continuity (ticks)	10	0%	0%	0%
	40	0%	0%	0%
	70	0%	12%	18%
	100	6%	98%	100%
	130	81%	100%	100%

Figure 21: new path direction probability

Also the average spread in figure 22 has a very sensitive reaction in this parameter area. The combination intensity of 5.5 and continuity of 10 has no valuable effect on the spread, but the spread changes with an increasing continuity to the direction of the forced norm.

		intensity (int)		
		5,5	6	6,5
continuity (ticks)	10	1,00	0,91	0,85
	40	0,92	0,48	0,42
	70	0,64	0,11	0,07
	100	0,25	-0,18	-0,21
	130	-0,07	-0,39	-0,41

Figure 22: Average spread, calculated at the last external impact tick

Conclusion

The aim of this research was to examine the behavior of external influences of a path dependent hierarchical organization with the method of computer simulation. The basis of the simulation was the M1 model from Petermann (2012) that simulates a path dependent hierarchical organization. The model was enhanced to simulate an external impact in form of continuity and intensity which were combined and incorporated into

the M1 model. For the first simulation a parameter setup that seemed to be valid during the implementation of the external impact was used. With the first results multiple questions arose and two more simulations with adjusted parameter setups were executed. To get an overview of the three simulations figure 23 shows an interesting chart spread versus intensity.

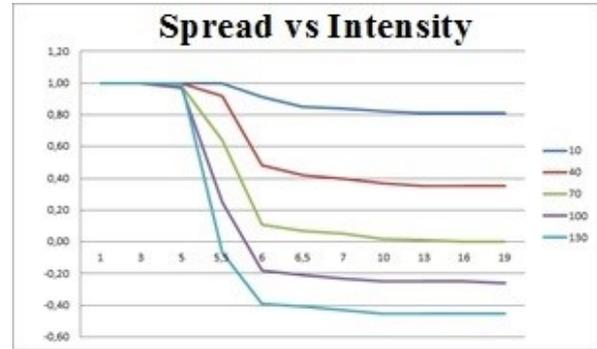


Figure 23: Spread vs intensity. X-axis: intensity. Y-axis: spread. Legend: Continuities from 10-130

To see a reaction on the system a critical intensity is needed. An intensity of 3 and less has no effect on the spread. The system first starts to react at the combination of intensity 3 and continuity of 100. As described in the third simulation, with an intensity of 5 the spread changes dramatically, but the intensive change stops immediately at the intensity of 6 and over. In this intensity field the external impact must last for a defined continuity to adopt a new norm in the whole company. The defined leadership impact of 1 concludes that with an intensity of 5, which is the minimum value to change the spread, the external impact needs to be five times stronger than the leadership impact. To clarify this, further research might show results.

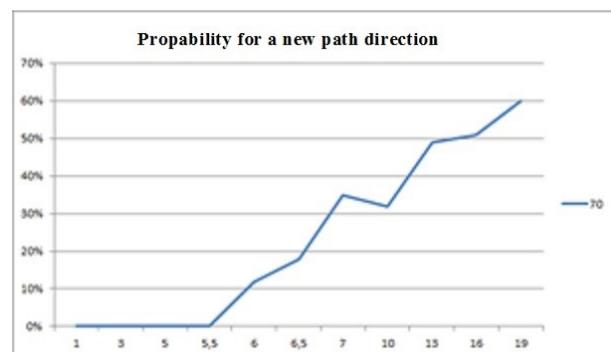


Figure 24: Probability for a new path direction with a continuity of 70. X-axis: intensity, y-axis: new path direction probability. Legend: continuity of 70

Another interesting result of this research is the new path direction probability. It was not in scope at the beginning of this research, but we figured out that we discovered an interesting system behavior at the continuity of 70 that leads the spread to 0 with

intensities of 13 and more. This probability increases more and more, the higher the intensity becomes. This unexpected system behavior should be investigated further in the future.

The next interesting point is the fact that path breaking does not necessarily lead to a new path direction. With a continuity of 40 and an intensity of 6.5 the path breaking probability is 100%, but the new path direction probability is 0%. The question that comes up here is: does it make sense to speak about breaking the path without actually changing the path? This might indicate the necessity to adapt the definition of path breaking in this context.

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.: 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

Prof. Dr. Arne Petermann is Professor of Management at the Berufsakademie für Gesundheits- und Sozialwesen Saarland (BAGSS). He is also Director of the Institute for Management and Quality at BAGSS. As a visiting scholar, he was teaching organization science and scientific simulation methods in the PhD-program at the School for Business and Economics at Freie Universität Berlin from 2007 to 2014. He is founder and CEO of Linara GmbH, an international HR agency specialized in the transnational European healthcare sector. He is also founder and CEO of 3P Projects GmbH, a consulting and executive education company based in Berlin, Germany. 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, MBA was born in Rheinfelden, Germany and studied electrical engineering at the University of Applied Sciences in St. Augustin, Germany and Business Administration Berlin University of Professional Studies. 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.

Applied Modelling And Simulation

REACHABILITY OF FRACTIONAL CONTINUOUS-TIME LINEAR SYSTEMS USING THE CAPUTO-FABRIZIO DERIVATIVE

Tadeusz Kaczorek
 Białystok University of Technology
 Faculty of Electrical Engineering
 Wiejska 45D, 15-351 Białystok
 E-mail: kaczorek@isep.pw.edu.pl

KEYWORDS

Fractional, continuous-time, linear, system, Caputo-Fabrizio definition, reachability.

ABSTRACT

The Caputo-Fabrizio definition of the fractional derivative is applied to analysis of the positivity and reachability of continuous-time linear systems. Necessary and sufficient conditions for the reachability of standard and positive fractional continuous-time linear systems are established.

INTRODUCTION

A dynamical system is called positive if its trajectory starting from any nonnegative initial condition state remains forever in the positive orthant for all nonnegative inputs. An overview of state of the art in positive system theory is given in the monographs (Farina and Rinaldi 2000; Kaczorek 2001) and in the papers (Kaczorek 1997, 1998, 2011b, 2014a, 2014b, 2015b). Models having positive behavior can be found in engineering, economics, social sciences, biology and medicine, etc.

The positive standard and descriptor systems and their stability have been analyzed in (Kaczorek 1997, 1998, 2001, 2011b, 2014b, 2015b). The positive linear systems with different fractional orders have been addressed in (Kaczorek 2011b, 2012) and the descriptor discrete-time linear systems in (Kaczorek 1998). Descriptor positive discrete-time and continuous-time nonlinear systems have been analyzed in (Kaczorek 2014a) and the positivity and linearization of nonlinear discrete-time systems by state-feedbacks in (Kaczorek 2014b). New stability tests of positive standard and fractional linear systems have been investigated in (Kaczorek 2011a). The stability and robust stabilization of discrete-time switched systems have been analyzed in (Zhang, Xie, Zhang and Wang 2014; Zhang, Han, Wu and Hung 2014). Minimum energy control of 2D systems in Hilbert spaces has been analyzed in (Klamka 1983). Controllability of dynamical systems has been investigated in (Kalman 1960; Klamka 1991, 1997, 1998).

Recently a new definition of the fractional derivative without singular kernel has been proposed in (Caputo and Fabrizio 2015; Losada and Nieto 2015).

In this paper the Caputo-Fabrizio definition of the fractional derivative will be applied to analysis of the reachability of the standard and positive linear systems. The paper is organized as follows. In section 2 necessary and sufficient conditions for the reachability of fractional standard continuous-time linear systems are established. Necessary and sufficient conditions for the positivity of the fractional systems and sufficient conditions for the reachability of the positive systems are proposed in section 3. Concluding remarks are given in section 4.

The following notation will be used: \mathfrak{R} - the set of real numbers, $\mathfrak{R}^{n \times m}$ - the set of $n \times m$ real matrices, $\mathfrak{R}_+^{n \times m}$ - the set of $n \times m$ matrices with nonnegative entries and $\mathfrak{R}_+^n = \mathfrak{R}_+^{n \times 1}$, M_n - the set of $n \times n$ Metzler matrices, I_n - the $n \times n$ identity matrix.

REACHABILITY OF STANDARD FRACTIONAL SYSTEMS

The Caputo-Fabrizio definition of fractional derivative of order α of the function $f(t)$ for $0 < \alpha < 1$ has the form (Caputo and Fabrizio 2015; Losada and Nieto 2015)

$${}^{CF}D^\alpha f(t) = \frac{1}{1-\alpha} \int_0^t \exp\left(-\frac{\alpha}{1-\alpha}(t-\tau)\right) \dot{f}(\tau) d\tau, \quad (1)$$

$$\dot{f}(\tau) = \frac{df(\tau)}{d\tau}, \quad t \geq 0.$$

Consider the fractional differential state equations

$${}^{CF}D^\alpha x(t) = \frac{d^\alpha x(t)}{dt^\alpha} = Ax(t) + Bu(t), \quad 0 < \alpha < 1, \quad (2a)$$

$$y(t) = Cx(t) + Du(t), \quad (2b)$$

where $x(t) \in \mathfrak{R}^n$, $u(t) \in \mathfrak{R}^m$, $y(t) \in \mathfrak{R}^p$ are the state, input and output vectors and $A \in \mathfrak{R}^{n \times n}$, $B \in \mathfrak{R}^{n \times m}$, $C \in \mathfrak{R}^{p \times n}$, $D \in \mathfrak{R}^{p \times m}$.

Theorem 1. The solution $x(t)$ of the equation (2a) for a given initial condition $x(0) = x_0$ and input $u(t)$ has the form

$$x(t) = e^{\hat{A}t}(\hat{x}_0 + \hat{B}u_0) + \int_0^t e^{\hat{A}(t-\tau)} \hat{B}[\beta u(\tau) + \dot{u}(\tau)]d\tau, \quad (3a)$$

where

$$\begin{aligned} \hat{A} &= \alpha[I_n - (1-\alpha)A]^{-1}A, \\ \hat{B} &= [I_n - (1-\alpha)A]^{-1}(1-\alpha)B, \quad \beta = \frac{\alpha}{1-\alpha}, \\ \hat{x}_0 &= [I_n - (1-\alpha)A]^{-1}x_0, \quad e^{\hat{A}t} = \mathcal{L}^{-1}\{I_n s - \hat{A}\}^{-1}, \\ \dot{u}(\tau) &= \frac{du(\tau)}{d\tau}, \quad u(0) = u_0. \end{aligned} \quad (3b)$$

Proof. The proof is given in (Kaczorek 2015a).

Definition 1. A state $x_f \in \mathfrak{R}^n$ of the standard system (2) is called reachable in time $t \in [0, t_f]$ if there exists an input $u(t) \in \mathfrak{R}^m$ for $t \in [0, t_f]$ which steers the state of the system from zero initial condition $x_0 = 0$ to the final state $x_f = x(t_f)$. If every state $x_f \in \mathfrak{R}^n$ is reachable in time $t \in [0, t_f]$ then the system is called reachable in time $t \in [0, t_f]$. The system (2) is called reachable if for every $x_f \in \mathfrak{R}^n$ there exists t_f and an input $u(t) \in \mathfrak{R}^m$ for $t \in [0, t_f]$ which steers the state of the system from $x_0 = 0$ to x_f .

Theorem 2. The standard fractional system (2) is reachable in time $t \in [0, t_f]$ if and only if the matrix

$$R_f = R(t_f) = \int_0^{t_f} e^{\hat{A}t} \hat{B} \hat{B}^T e^{\hat{A}^T t} dt \quad (4)$$

is invertible.

The input which steers the state of the system from $x_0 = 0$ to x_f is given by

$$u(t) = \int_0^t e^{-\beta\tau} \hat{B}^T e^{\hat{A}^T(t_f-\tau)} d\mathfrak{T}R_f^{-1} x_f, \quad t \in [0, t_f] \quad (5)$$

and $u_0 = u(0) = 0$.

Proof. Substituting

$$\bar{u}(t) = \beta u(t) + \dot{u}(t) \quad (6)$$

into (3a) for $x_0 = 0$, $u_0 = 0$ we obtain

$$x(t) = \int_0^t e^{\hat{A}(t-\tau)} \hat{B} \bar{u}(\tau) d\tau. \quad (7)$$

The solution of the differential equation (6) for $u_0 = u(0) = 0$ has the form

$$u(t) = \int_0^t e^{-\beta\tau} \bar{u}(t-\tau) d\tau. \quad (8)$$

To show that the input

$$\bar{u}(t) = \hat{B}^T e^{\hat{A}(t_f-t)} R_f^{-1} x_f, \quad t \in [0, t_f] \quad (9)$$

steers the state from $x_0 = 0$ to x_f in time $t \in [0, t_f]$ we substitute (9) into (7) and we obtain

$$\begin{aligned} x(t_f) &= \int_0^{t_f} e^{\hat{A}(t_f-\tau)} \hat{B} \hat{B}^T e^{\hat{A}^T(t_f-\tau)} d\mathfrak{T}R_f^{-1} x_f \\ &= R_f R_f^{-1} x_f = x_f. \end{aligned} \quad (10)$$

Substituting (9) into (8) we obtain (5). \square

From Theorem 1 and its proof follows the corollary.

Corollary 1. The fractional system (2) is reachable in time $t \in [0, t_f]$ if and only if the fractional system

$$\frac{d^\alpha x(t)}{dt^\alpha} = \hat{A}x(t) + \hat{B}u(t) \quad (11)$$

is reachable in time $t \in [0, t_f]$.

The input $\bar{u}(t)$ steers the state $x(t)$ from $x_0 = 0$ to x_f in time $t \in [0, t_f]$ of the system (11) if and only if the input (8) steers the state from $x_0 = 0$ to x_f in time $t \in [0, t_f]$ of the system (2a).

Example 1. Consider the fractional system described by the equation (2a) with $\alpha = 0.5$, zero initial condition $x_0 = 0$, $u_0 = 0$ and

$$A = \begin{bmatrix} -1 & a \\ 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad a - \text{parameter}. \quad (12)$$

Compute the input $u(t)$ which steers the system from $x_0 = 0$ to $x_f = [1 \ 1]^T$ (T denotes transpose) in time $t \in [0, 1]$. Using (3b) and (12) we obtain

$$\begin{aligned} \hat{A} &= \alpha[I_2 - (1-\alpha)A]^{-1}A \\ &= 0.5 \begin{bmatrix} 1.5 & -0.5a \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} -1 & a \\ 0 & -2 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} -2 & a \\ 0 & -3 \end{bmatrix}, \end{aligned} \quad (13a)$$

$$\begin{aligned}\hat{B} &= [I_2 - (1-\alpha)A]^{-1}(1-\alpha)B \\ &= \begin{bmatrix} 1.5 & -0.5a \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} a+4 \\ 3 \end{bmatrix}\end{aligned}\quad (13b)$$

Taking into account that the eigenvalues of the matrix (13a) are $\lambda_1 = -\frac{1}{2}$, $\lambda_2 = -\frac{1}{3}$ and using the Sylvester formula we obtain

$$\begin{aligned}e^{\hat{A}t} &= \frac{\hat{A} - I_2 \lambda_2}{\lambda_1 - \lambda_2} e^{\lambda_1 t} + \frac{\hat{A} - I_2 \lambda_1}{\lambda_2 - \lambda_1} e^{\lambda_2 t} \\ &= \begin{bmatrix} 0 & -a \\ 0 & 1 \end{bmatrix} e^{\frac{1}{2}t} + \begin{bmatrix} 1 & a \\ 0 & 0 \end{bmatrix} e^{\frac{1}{3}t} \\ &= \begin{bmatrix} e^{\frac{1}{3}t} & a \left(e^{\frac{1}{3}t} - e^{\frac{1}{2}t} \right) \\ 0 & e^{\frac{1}{2}t} \end{bmatrix}.\end{aligned}\quad (14)$$

Using (4) for $t_f = 1$ and (14), (13b) we obtain

$$\begin{aligned}R_f &= \int_0^{t_f} e^{\hat{A}t} \hat{B} \hat{B}^T e^{\hat{A}^T t} dt = \int_0^1 (e^{\hat{A}t} \hat{B})(e^{\hat{A}t} \hat{B})^T dt \\ &= \int_0^1 \begin{bmatrix} a \left(\frac{2}{3} e^{\frac{1}{3}t} - \frac{1}{2} e^{\frac{1}{2}t} \right) + \frac{2}{3} e^{\frac{1}{3}t} \\ \frac{1}{2} e^{\frac{1}{2}t} \end{bmatrix} \\ &\times \begin{bmatrix} a \left(\frac{2}{3} e^{\frac{1}{3}t} - \frac{1}{2} e^{\frac{1}{2}t} \right) + \frac{2}{3} e^{\frac{1}{3}t} & \frac{1}{2} e^{\frac{1}{2}t} \end{bmatrix} dt \\ &= \int_0^1 \begin{bmatrix} \left[a \left(\frac{2}{3} e^{\frac{1}{3}t} - \frac{1}{2} e^{\frac{1}{2}t} \right) + \frac{2}{3} e^{\frac{1}{3}t} \right]^2 \\ \left[a \left(\frac{2}{3} e^{\frac{1}{3}t} - \frac{1}{2} e^{\frac{1}{2}t} \right) + \frac{2}{3} e^{\frac{1}{3}t} \right] \frac{1}{2} e^{\frac{1}{2}t} \\ \left[a \left(\frac{2}{3} e^{\frac{1}{3}t} - \frac{1}{2} e^{\frac{1}{2}t} \right) + \frac{2}{3} e^{\frac{1}{3}t} \right] \frac{1}{2} e^{\frac{1}{2}t} \\ \left[\frac{1}{2} e^{\frac{1}{2}t} \right]^2 \end{bmatrix} dt \\ &= \begin{bmatrix} 0.0301a^2 + 0.1965a + 0.3244 \\ 0.0681a + 0.2262 \\ 0.0681a + 0.2262 \\ 0.158 \end{bmatrix}.\end{aligned}\quad (15)$$

The matrix (15) is nonsingular since $\det R_f = 0.0001a^2 + 0.0002a + 0.0001 \neq 0$ for $a \neq -1$ and by Theorem 1 the fractional system with (12) is reachable in time $t \in [0,1]$ for $a \neq -1$.

The input steering the system from $x_0 = 0$ to $x_f = [1 \ 1]^T$ in time $t \in [0,1]$ is given by

$$\begin{aligned}u(t) &= \int_0^t e^{-\beta(t-\tau)} \hat{B}^T e^{\hat{A}^T(t_f-\tau)} d\tau R_f^{-1} x_f \\ &= e^{-t} \int_0^t e^{-\tau} \hat{B}^T e^{\hat{A}^T} e^{-\hat{A}^T \tau} d\tau R_f^{-1} x_f \\ &= e^{-t} \int_0^t e^{\tau} \frac{1}{6} [a+4 \ 3] \begin{bmatrix} 0.7165 & 0 \\ 0.11a & 0.6065 \end{bmatrix} \\ &\times \begin{bmatrix} e^{\frac{1}{3}\tau} & 0 \\ a \begin{pmatrix} \frac{1}{3}\tau & \frac{1}{2}\tau \end{pmatrix} & e^{\frac{1}{2}\tau} \end{bmatrix} d\tau \\ &\times \begin{bmatrix} 0.0301a^2 + 0.1965a + 0.3244 \\ 0.0681a + 0.2262 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= e^{-t} \int_0^t e^{\tau} \begin{bmatrix} e^{\frac{1}{3}\tau} (0.4777a + 0.4777) - 0.333e^{\frac{1}{2}\tau} \\ 0.3033e^{\frac{1}{2}\tau} \end{bmatrix} d\tau \\ &\times \begin{bmatrix} -0.0681a - 0.0682 \\ 0.0001a^2 + 0.0002a + 0.0001 \\ 0.0301a^2 + 0.1284a + 0.0982 \\ 0.0001a^2 + 0.0002a + 0.0001 \end{bmatrix}.\end{aligned}\quad (16)$$

For example for $a = 1$ we obtain

$$\begin{aligned}u(t) &= e^{-t} \int_0^t e^{\tau} \begin{bmatrix} e^{\frac{1}{3}\tau} (0.4777a + 0.4777) - 0.303e^{\frac{1}{2}\tau} \\ 0.3033e^{\frac{1}{2}\tau} \end{bmatrix} d\tau \begin{bmatrix} -340.75 \\ 641.75 \end{bmatrix} \\ &= -244.1644e^{\frac{1}{3}t} + 198.6615e^{\frac{1}{2}t} + 45.2029e^{-t}.\end{aligned}\quad (17)$$

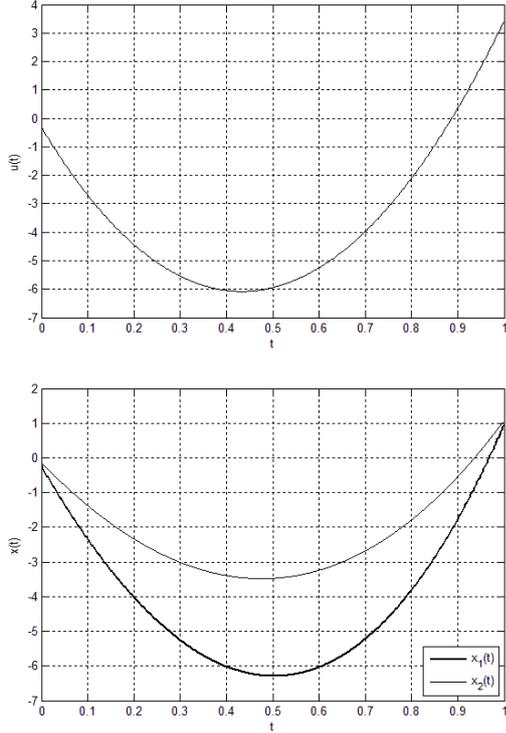


Figure 1: Input signal and state vector for $t_f = 1[s]$.

REACHABILITY OF POSITIVE FRACTIONAL SYSTEMS

Consider the fractional system (2).

Definition 2. The fractional system (2) is called (internally) positive if the state vector $x(t) \in \mathfrak{R}_+^n$ and the output vector $y(t) \in \mathfrak{R}_+^p$, $t \geq 0$ for all initial conditions $x_0 \in \mathfrak{R}_+^n$ and all inputs $u(t) \in \mathfrak{R}_+^m$, $\dot{u}(t) \in \mathfrak{R}_+^m$, $t \geq 0$.

Definition 3. A real matrix $A = [a_{ij}] \in \mathfrak{R}^{n \times n}$ is called Metzler matrix if its off-diagonal entries are nonnegative, i.e. $a_{ij} \geq 0$ for $i \neq j$; $i, j = 1, \dots, n$.

Lemma 1. Let $\hat{A} \in M_n$ and $0 < \alpha < 1$. Then

$$e^{\hat{A}t} \in \mathfrak{R}_+^{n \times n} \text{ for } t \geq 0. \quad (18)$$

Proof. The proof is similar to the one given in (Kaczorek 2001).

Theorem 3. The fractional system (2) is positive if and only if

$$\hat{A} \in M_n, \hat{B} \in \mathfrak{R}_+^{n \times m}, C \in \mathfrak{R}_+^{p \times n}, D \in \mathfrak{R}_+^{p \times m}. \quad (19)$$

Proof. Sufficiency. If $\hat{A} \in M_n$ and $\hat{B} \in \mathfrak{R}_+^{n \times m}$ then from (3) we have $x(t) \in \mathfrak{R}_+^n$, $t \geq 0$ since by Lemma 1 $e^{\hat{A}t} \in \mathfrak{R}_+^{n \times n}$ and $x_0 \in \mathfrak{R}_+^n$, $u(t) \in \mathfrak{R}_+^m$, $\dot{u}(t) \in \mathfrak{R}_+^m$, $t \geq 0$.

Necessity. Let $u(t) = 0$, $t \geq 0$ and $x_0 = e_i$ (i -th column of the identity matrix I_n). The trajectory does not leave the orthant \mathfrak{R}_+^n only if ${}^{CF}D^\alpha x(0) = \hat{A}e_i \geq 0$ what implies $\hat{a}_{ij} \geq 0$ for $i \neq j$, $i, j = 1, \dots, n$ and $\hat{A} \in M_n$. If $x_0 = 0$ then ${}^{CF}D^\alpha x(0) = Bu(0) \geq 0$ and this implies $B \in \mathfrak{R}_+^{n \times m}$ since $u(0) \in \mathfrak{R}_+^m$ is arbitrary. From (2b) for $u(t) = 0$, $t \geq 0$ we have $y(0) = Cx(0)$ and $C \in \mathfrak{R}_+^{p \times n}$ since $x(0) = x_0 \in \mathfrak{R}_+^n$ is arbitrary. Assuming $x_0 = 0$ from (2b) we have $y(0) = Du(0)$ and $D \in \mathfrak{R}_+^{p \times m}$ since $u(0) \in \mathfrak{R}_+^m$ is arbitrary. \square

Lemma 2. If λ_k , $k = 1, \dots, n$ are the eigenvalues of the matrix A then the eigenvalues of the matrix $\hat{A} = \alpha[I_n - (1 - \alpha)A]^{-1}A$ are given by

$$\hat{\lambda}_k = \alpha[1 - (1 - \alpha)\lambda_k]^{-1}\lambda_k. \quad (20)$$

Proof. It is well-known (Gantmacher 1959) that if $f(\lambda_k)$ is well-defined on the spectrum λ_k , $k = 1, \dots, n$ of the matrix A then the eigenvalues of the matrix $f(A)$ are given by $f(\lambda_k)$, $k = 1, \dots, n$. In this case $f(A) = \alpha[I_n - (1 - \alpha)A]^{-1}A$. \square

Lemma 3. The matrix $\bar{A} = (1 - \alpha)A \in \mathfrak{R}^{n \times n}$ for $0 < \alpha < 1$ is asymptotically stable if and only if the matrix A is asymptotically stable.

Proof. The eigenvalues $\bar{\lambda}_k$, $k = 1, \dots, n$ of the matrix \bar{A} are related with the eigenvalues λ_k , $k = 1, \dots, n$ of the matrix A by

$$\bar{\lambda}_k = (1 - \alpha)\lambda_k, \quad k = 1, \dots, n. \quad (21)$$

since the characteristic polynomials of the matrices are related by the equality

$$\begin{aligned} \det[I_n \bar{\lambda}_k - \bar{A}] &= \det[I_n \bar{\lambda}_k - (1 - \alpha)A] \\ &= (1 - \alpha)^n \det\left[I_n \frac{\bar{\lambda}_k}{1 - \alpha} - A\right] \\ &= (1 - \alpha)^n \det[I_n \lambda_k - A]. \end{aligned} \quad (22)$$

Therefore, from (21) it follows that $\text{Re } \bar{\lambda}_k < 0$, $k = 1, \dots, n$ if and only if $\text{Re } \lambda_k < 0$, $k = 1, \dots, n$. \square

Lemma 4. The matrix

$$\hat{A} = \alpha[I_n - (1 - \alpha)A]^{-1}A \in M_n \quad (23)$$

is asymptotically stable if and only if the eigenvalues $\lambda_k = -\alpha_k + j\beta_k$, $k = 1, \dots, n$ of the matrix A satisfy the condition

$$[1 + (1 - \alpha)\alpha_k] \alpha_k + (1 - \alpha)\beta_k^2 = n(k) > 0.$$

Proof. From (20) for $\hat{\lambda}_k = -\hat{\alpha}_k + j\hat{\beta}_k$ and $\lambda_k = -\alpha_k + j\beta_k$, $k = 1, \dots, n$ we have

$$\begin{aligned} \hat{\lambda}_k &= -\hat{\alpha}_k + j\hat{\beta}_k = \alpha[1 - (1 - \alpha)\lambda_k]^{-1} \lambda_k \\ &= \alpha[1 - (1 - \alpha)(-\alpha_k + j\beta_k)]^{-1} (-\alpha_k + j\beta_k) \\ &= \alpha \frac{1 + (1 - \alpha)\alpha_k + j(1 - \alpha)\beta_k}{[1 + (1 - \alpha)\alpha_k]^2 + [(1 - \alpha)\beta_k]^2} (-\alpha_k + j\beta_k) \quad (24) \\ &= \alpha \left(\frac{-[1 + (1 - \alpha)\alpha_k] \alpha_k - (1 - \alpha)\beta_k^2}{[1 + (1 - \alpha)\alpha_k]^2 + [(1 - \alpha)\beta_k]^2} \right. \\ &\quad \left. + j \frac{[1 + (1 - \alpha)\alpha_k] \beta_k - (1 - \alpha)\alpha_k \beta_k}{[1 + (1 - \alpha)\alpha_k]^2 + [(1 - \alpha)\beta_k]^2} \right) \end{aligned}$$

and

$$\begin{aligned} \hat{\alpha}_k &= \alpha \left(\frac{[1 + (1 - \alpha)\alpha_k] \alpha_k + (1 - \alpha)\beta_k^2}{[1 + (1 - \alpha)\alpha_k]^2 + [(1 - \alpha)\beta_k]^2} \right) \\ &= \alpha \frac{n(k)}{d(k)}, \quad k = 1, \dots, n. \end{aligned} \quad (25)$$

From (25) it follows that $\hat{\alpha}_k > 0$, $k = 1, \dots, n$ if and only if $n(k) > 0$, $k = 1, \dots, n$. \square

Lemma 5. The matrices

$$\begin{aligned} \hat{A} &= \alpha[I_n - (1 - \alpha)A]^{-1} A \in M_n, \\ \hat{B} &= [I_n - (1 - \alpha)A]^{-1} (1 - \alpha)B \in \mathfrak{R}_+^{n \times m} \end{aligned} \quad (26)$$

if $A \in M_n$ is asymptotically stable and $B \in \mathfrak{R}_+^{n \times m}$.

Proof. The matrix $[I_n - (1 - \alpha)A]^{-1} \in \mathfrak{R}_+^{n \times n}$ if the matrix $A \in M_n$ is asymptotically stable (Kaczorek 2001). Therefore, by Lemma 3 and $(1 - \alpha)B \in \mathfrak{R}_+^{n \times m}$ for $0 < \alpha < 1$ (25) holds if $A \in M_n$ is asymptotically stable. \square

From Lemma 4 and Theorem 3 we have the following.

Theorem 4. The fractional system (2) is positive if $A \in M_n$ is asymptotically stable and $B \in \mathfrak{R}_+^{n \times m}$, $C \in \mathfrak{R}_+^{p \times n}$, $D \in \mathfrak{R}_+^{p \times m}$.

Definition 4. A state $x_f \in \mathfrak{R}_+^n$ of the positive system (2) is called reachable in time $t \in [0, t_f]$ if there exists an input $u(t) \in \mathfrak{R}_+^m$ for $t \in [0, t_f]$ which steers the state of the system from zero initial condition $x_0 = 0$ to the final state $x_f \in \mathfrak{R}_+^n$. If every state $x_f \in \mathfrak{R}_+^n$ is reachable in time $t \in [0, t_f]$ then the system is called reachable in

time $t \in [0, t_f]$. The positive system (2) is called reachable if for every $x_f \in \mathfrak{R}_+^n$ there exists t_f and an input $u(t) \in \mathfrak{R}_+^m$ for $t \in [0, t_f]$ which steers the state of the system from $x_0 = 0$ to x_f .

Definition 5. A matrix $A \in \mathfrak{R}^{n \times n}$ is called monomial if in each row and in each column only one entry is positive and the remaining entries are zero.

Theorem 5. The positive fractional system (2) is reachable in time $t \in [0, t_f]$ if the matrix

$$R_f = R(t_f) = \int_0^{t_f} e^{\hat{A}t} \hat{B} \hat{B}^T e^{\hat{A}^T t} dt \quad (27)$$

is monomial.

The input which steers the state of the system from $x_0 = 0$ to x_f is given by

$$u(t) = \int_0^t e^{-\beta\tau} \hat{B}^T e^{\hat{A}^T(t_f - \tau)} d\tau R_f^{-1} x_f \quad (28)$$

Proof. It is well-known (Kaczorek 2001) that $R_f^{-1} \in \mathfrak{R}_+^{n \times n}$ if and only if the matrix $R_f \in \mathfrak{R}_+^{n \times n}$ is monomial. In a similar way as in proof of Theorem 1 it can be shown that the input (28) steers the state of positive system from $x_0 = 0$ to $x_f \in \mathfrak{R}_+^n$ in time $t \in [0, t_f]$. From (28) it follows that $u(t) \in \mathfrak{R}_+^m$ since

$$e^{-\beta t} > 0 \quad \text{for} \quad \beta = \frac{\alpha}{1 - \alpha} > 0, \quad 0 < \alpha < 1,$$

$$\hat{B}^T e^{\hat{A}^T(t_f - \tau)} \in \mathfrak{R}_+^{m \times n} \text{ and } R_f^{-1} x_f \in \mathfrak{R}_+^n. \quad \square$$

Example 2. (Continuation of Example 1) Note that the matrix R_f given by (15) is monomial only for $a = -3.3216$. Therefore, we cannot say anything about the reachability of the positive system with (12) in time $t \in [0, 1]$ for $a \geq 0$.

CONCLUDING REMARKS

The Caputo-Fabrizio definition of the fractional derivative has been applied to analysis of the positivity and reachability of continuous-time linear systems. Necessary and sufficient conditions for the reachability of standard continuous-time linear systems have been established (Theorem 1). Necessary and sufficient conditions for the positivity of the fractional linear systems have been given (Theorems 3 and 4). Sufficient conditions for the reachability of the fractional positive linear systems have been also established (Theorem 5). The considerations are illustrated by numerical examples of standard and positive fractional linear systems.

The considerations can be extended to continuous-discrete linear systems.

ACKNOWLEDGEMENT

This work was supported by National Science Centre in Poland under work No. 2014/13/B/ST7/03467.

REFERENCES

- Caputo M. and Fabrizio M. 2015. "A New Definition of Fractional Derivative without Singular Kernel". *Progr. Fract. Differ. Appl.*, Vol.1, No.2, 1-13.
- Farina L. and Rinaldi S. 2000. "Positive Linear Systems". J. Wiley, New York.
- Gantmacher F.R. 1959. "The Theory of Matrices". Chelsea Pub. Comp., London.
- Kaczorek T. 1997. "Positive singular discrete time linear systems". *Bull. Pol. Acad. Techn. Sci.*, Vol.45, No.4, 619-631.
- Kaczorek T. 1998. "Positive descriptor discrete-time linear systems". *Problems of Nonlinear Analysis in Engineering Systems*, Vol.1, No.7, 38-54.
- Kaczorek T. 2001. "Positive 1D and 2D Systems". Springer-Verlag, London.
- Kaczorek T. 2011a. "New stability tests of positive standard and fractional linear systems". *Circuits and Systems*, Vol.2, No.4, 261-268.
- Kaczorek T. 2011b. "Positive linear systems consisting of n subsystems with different fractional orders". *IEEE Trans. Circuits and Systems*, Vol.58, No.6, 1203-1210.
- Kaczorek T. 2012. "Selected Problems of Fractional Systems Theory". Springer-Verlag, Berlin.
- Kaczorek T. 2014a. "Descriptor positive discrete-time and continuous-time nonlinear systems". *Proc. of SPIE*, Vol.9290, doi:10.1117/12.2074558.
- Kaczorek T. 2014b. "Positivity and linearization of a class of nonlinear discrete-time systems by state feedbacks". *Logistyka*, Vol.6, 5078-5083.
- Kaczorek T. 2015a. "Analysis of positive and stable fractional continuous-time linear systems by the use of Caputo-Fabrizio derivative". Submitted to *IEEE Trans. Circuits and Systems*.
- Kaczorek T. 2015b. "Positivity and stability of discrete-time nonlinear systems". *Proc. of CYBCONF*.
- Kalman R.E. 1960. "On the general theory of control systems". *Proc. of the first Intern. Congress on Automatic Control*, London, 481-493.
- Klamka J. 1983. "Minimum energy control of 2D systems in Hilbert spaces". *Systems Science*, Vol.9, No.1-2, 33-42.
- Klamka J. 1991. "Controllability of Dynamical Systems". Kluwer Academic Press, Dordrecht.
- Klamka J. 1997. "Controllability of 2-D systems: a survey". *Int. J. Appl. Math. Comput. Sci.*, Vol.7, No.4, 101-120.
- Klamka J. 1998. "Constrained controllability of positive 2-D systems". *Bull. Pol. Acad. Techn. Sci.*, Vol.61, No.1, 95-104.
- Losada J. and Nieto J. 2015. "Properties of a new fractional derivative without singular kernel". *Progr. Fract. Differ. Appl.*, Vol.1, No.2, 87-92.
- Zhang H.; Xie D.; Zhang H.; and Wang G. 2014. "Stability analysis for discrete-time switched systems with unstable subsystems by a mode-dependent average dwell time approach". *ISA Transactions*, Vol.53, 1081-1086.
- Zhang H.; Han Z.; Wu H.; and Hung J. 2014. "Robust stabilization of discrete-time positive switched systems with uncertainties and average dwell time switching". *Circuits Syst. Signal Process.*, Vol.33, 71-95.



AUTHOR BIOGRAPHIES

TADEUSZ KACZOREK received the M.Sc., Ph.D. and D.Sc. degrees in electrical engineering from the Warsaw

University of Technology in 1956, 1962 and 1964, respectively. In the years 1968–69 he was the dean of the Electrical Engineering Faculty, and in the period of 1970–73 he was a deputy rector of the Warsaw University of Technology. In 1971 he became a professor and in 1974 a full professor at the same university. Since 2003 he has been a professor at the Białystok University of Technology. In 1986 he was elected a corresponding member and in 1996 a full member of the Polish Academy of Sciences. In the years 1988–1991 he was the director of the Research Centre of the Polish Academy of Sciences in Rome. In 2004 he was elected an honorary member of the Hungarian Academy of Sciences. He was granted honorary doctorates by 13 universities. His research interests cover systems theory, especially singular multidimensional systems, positive multidimensional systems, singular positive 1D and 2D systems, as well as positive fractional 1D and 2D systems. He initiated research in the field of singular 2D, positive 2D and positive fractional linear systems. He published 28 books (8 in English) and over 1000 scientific papers. He also supervised 69 Ph.D. theses. He is the editor-in-chief of the *Bulletin of the Polish Academy of Sciences: Technical Sciences* and a member of editorial boards of ten international journals.

SIMULATION IMPROVES OPERATIONS AT A SPECIALIZED TAKEOUT RESTAURANT

Sapthagirishwaran Thennal Sivaramakrishnan, Shanmugasundaram Chandrasekaran, Jennifer Dhanapal,
Paul Ajaydivyan Jeya Sekar, Edward J. Williams

Decision Science, College of Business
University of Michigan – Dearborn
Fairlane Center South, 19000 Hubbard Drive
Dearborn MI 48126, USA

KEYWORDS

Restaurant operations, discrete-event process simulation

ABSTRACT

For more than half a century now, discrete-event process simulation has repeatedly proved itself a powerful analytical tool for improving many types of commercial and industrial processes. This analytical power is especially highly valued when the operational complexity and/or stochastic variability of the process exceeds the ability of closed-form equations to model it. Historically, simulation first proved its worth, and was most extensively used, in the analysis and improvement of manufacturing operations. More recently, the use of simulation has expanded vigorously and broadly to include warehousing operations, the delivery of health care (hospitals and clinics), transportation services (airlines, railroads, and bus lines), and the hospitality industry (amusement parks, hotels, restaurants, and cruise ships).

In the successful simulation application described in this paper, simulation was used to model, analyze, and improve the staffing levels and operational procedures of a restaurant – unusually, a restaurant which provides *only* take-out services, with (by business choice) *no* “dine-in” capacity. The simulation analysis showed the most effective path to correction of insufficient capacity, distressingly long waiting times, and consequent lost sales and revenue.

INTRODUCTION

Historically, the first vigorous commercial uses of discrete-event process simulation were in the manufacturing sectors of the economy processes (Law and McComas 1999). More recently, and now very aggressively and successfully, simulation analysts and industrial engineers have expanded its use to warehousing, the delivery of health care, the operation of transportation networks, and the delivery of consumer services. These consumer services are wide-ranging, including the operation of retail stores, banks, hotels, and restaurants. Hotels and restaurants are two of the key sectors in the “hospitality industry;” (Starks and Whyte 1998) provides a tutorial on the use of

simulation in this industry, with emphasis on the study of fast-food restaurants. At a more detailed level, (Brann and Kulick 2002) describes the simulation of restaurant operations, and (Curin et al. 2005) describes the successful use of simulation to reduce service times at a busy fast-food university campus restaurant. The focus of analysis in the present study was likewise a restaurant, specifically one providing take-away dinners but offering no dine-in services. The restaurant owners and managers embraced the use of simulation to explore solutions to long-standing problems of inadequate staffing, resulting long waiting times for order pick-up, customer dissatisfaction, and ultimately lost sales.

This paper is organized as follows: The next section provides a high-level description of the restaurant’s operations. The following two sections explain, in turn, the collection and analysis of the input data; and then the construction, verification, and validation of the simulation model. The next section describes the results and conclusions obtained by experimentation runs of the model. The final section presents overall conclusions, how the results specifically guided and helped the restaurateur, and indicated future work.

OVERVIEW OF THE RESTAURANT’S OPERATIONS

The restaurant in question, Veggie Delight, is located near the center of one of south India’s most populous and busy cities, Chennai. Its highly successful business model is to serve takeout only, during the dinner pickup hours (6pm to 10pm, although orders accepted before 10pm may perform be picked up shortly after that hour), and to serve only one famous south Indian dish, idli (south Indian rice cake) with chutney (a family of condiments – hence this is a vegetarian entrée) (Achaya 2012). No diners are served within the restaurant itself. All orders are taken by telephone; the typical customer telephones the dinner order from his or her workplace, then intending to drive by the restaurant on the way home and pick up the boxed dinner for leisurely home consumption. A surge of telephone calls begins immediately after the 6pm opening time, and continues unabated through the evening, as workers head home from often late working hours in nearby offices. Indeed, the restaurant

proprietors might say the business model has become *too* successful – leading to overload of the cooks and the kitchen capacity, hence long wait times, impatient customers, and hence complaints and lost sales. Many of the analytical challenges presented by this business context were strikingly similar to those presented by the simulation of an oil change center (Williams et al. 2005). In both contexts, the customer service was provided on a basis of “drive in, receive needed service (the provision of fresh oil or a boxed dinner), and drive out.” In both businesses, the most pressing problems related to staffing levels and the deployment of personnel. In the case of this restaurant, when the simulation study began, there were two workers – one dedicated to answering the telephone and one working to cook and package the idlies – and four telephone lines. And both simulation clients expressed the heartfelt concern “The line is so bad it’s out into the street.”

Confronting challenges such as these, the restaurant management sought to determine the potential improvements to key performance metrics which could be obtained by increasing the number of incoming telephone lines and/or the number of culinary workers. The key performance metrics were:

1. Rejected (“dropped”) incoming telephone orders per hour
2. Utilization of the kitchen workers
3. Customers’ time-in-system
4. Average number of waiting customers, and their average waiting time
5. Number of customers served

INPUT DATA AND ITS ANALYSIS

Since the restaurant is a specialized one which opens each day at 6pm (to cater to the dinner trade) and closes shortly after 10pm (when the last customer arriving before 10pm has received full service), data collection was not unduly burdensome in terms of time required. On multiple days, distributed among the days of the week, the modeling team collected – by direct observation – data pertaining to the number of incoming calls received *and answered* per hour, the time required to answer a call, the number of idlies ordered in a call, the time required to prepare idlies, and the time required to package them for handover to the arriving customer. It was readily observed that the packaging of cooked idlies was not a bottleneck of concern. The restaurant has telephone lines scrupulously kept available for incoming calls (employees are not allowed to use these lines to place an outbound call). The italicized words above (“and answered”) acknowledge an admitted deficiency of this data collection: If a would-be customer places a call when all deployed lines are already receiving orders, that customer will receive a busy signal and the failure of that call goes undetected.

These data were analyzed with the commonly used and reliable distribution-fitting software Stat::Fit®. Appropriate techniques of using a software package for this purpose appear in (Chung 2004). The Stat::Fit® software supports the chi-squared, Kolmogorov-Smirnov, and Anderson-Darling measures of fit quality, as documented in a description of this software which appears in (Leemis 2002).

Results of these analyses led to the following conclusions:

1. Call interarrival times (for calls successfully received) were exponentially distributed with mean six minutes.
2. Time taken to answer a customer call was uniformly distributed between two and two and a half minutes.
3. The time required by the kitchen to prepare an idli was triangularly distributed with minimum seven minutes, mode eight minutes, and maximum nine minutes.
4. The time required to pack one, two, or three idlies in a box (a box holds a maximum of three idlies) was triangularly distributed with minimum 30 seconds, mode 48 seconds, and maximum 60 seconds.

The number of idlies ordered was distributed as follows:

Table 1. Probabilities of Ordered Quantities

Number of Idlies Ordered	Probability
1	0.20
2	0.30
3	0.30
4	0.15
5	0.05

The number of idlies ordered was found independent of time of evening and independent of day of week. Therefore, in the model, a customized discrete distribution was defined to represent these probabilities.

No significant downtimes of equipment, nor absenteeism problems of workers, were noticed during the data collection period. Therefore, no such downtimes or personnel shortages were incorporated into the simulation model. Furthermore, the success the staffing increase (two more workers) achieved when implemented convinced the restaurateur to use simulation analysis again to evaluate contingency plans addressing the quantitative effects of unexpected absenteeism, equipment failure, or electrical failure (this last a highly pertinent concern, especially during the rainy season when frequent power outages are quite likely).

MODEL CONSTRUCTION, VERIFICATION, AND VALIDATION

The simulation software tool Simio®, well documented in (Kelton, Smith, and Sturrock 2013), (Joines and Roberts 2015), and (Thiesing and Pegden 2015), was used for this project. Simio® constructs such as Servers (representing the cooking or the packing stations), Sources (representing arrival of either incoming calls or of customers “physically” arriving to pick up orders), Sinks (representing calls or customers leaving the system), and Resources (representing cooks or packers) were well-suited to model the process. Entities moving through this model represented incoming calls, arriving customers, or idlies being cooked, packaged, and delivered to customers. Model logic using entity attributes (which Simio® calls “states”) ensured the delivery of an order of idlies to the customer who placed the order by telephone. Commendably, Simio® supports the development of model logic via a “drag-and-drop” flowchart-construction interface in lieu of the writing of code in a software language (typically akin to Visual Basic for Applications). This support helps make simulation analysis to graduate-level business students, who – in the United States – typically do *not* have a background in computer coding in a language such as VBA or C++. An example showing the logic executed upon the acceptance of an order appears as Figure 1 in the Appendix. Similar logic ensures that no incoming calls can enter the system after 10pm – but orders received previously will move through the system until fulfillment and the “last order out” will “close” the restaurant.

Various techniques (Sargent 2015) were used to verify and validate the model. These techniques included:

1. Sending *one* arriving order (followed by *one* arriving customer to pick it up) through the system and tracing it step-by-step via the animation. The order and customer entities proceeded through the model correctly, including successful matching of the Customer entity with the Order entity, at the “Delivery” point (which Simio® calls a “Combiner”).
2. Directional testing – for example, increasing or decreasing the frequency of incoming calls beyond plausible limits and checking that performance metrics such as resource utilizations increased or decreased as expected. For example, at the “Kitchen” (a “Server” in Simio® nomenclature) queues grew steadily as incoming call rates increased – until the utilization of call lines reached 100%.
3. Undertaking structured walkthroughs of process logic within the modeling team.
4. Checking (via display of the state variable “number of entities in system” on the animation) that this number decreased monotonically after 10pm.

5. Reconciling the total number of idlies cooked and delivered with the number of customers served and the distribution of “idlies ordered per customer” as specified in Table 1. For example, as test run specifying that exactly 100 customers would enter the system should lead to an expectation of selling 255 ($20*1 + 30*2 + 30*3 + 15*4 + 5*5$) idlies; several such test runs all predicted sales of between 241 and 266 idlies).

At the conclusion of these verification and validation endeavors, and the routine correction of mistakes exposed by these efforts, the simulation model was deemed valid and acknowledged as credible by the restaurant proprietor. Specifically, all performance metrics of high interest (number of customers served, number of idlies cooked and delivered, time-in-system of incoming calls, and utilizations of cooks and packers) matched current observations in the restaurant to within 4%.

RESULTS OF THE SIMULATION MODEL

Since the restaurant opens afresh every day at 6pm and runs until “empty and after 10pm,” simulation runs were terminating, not steady-state, and hence needed zero warm-up time. Each of sixteen scenarios (one to four telephone lines and one to four workers) was run for twenty replications.

Originally, the restaurant had four telephone lines and two workers. After examining the simulation results, the proprietor decided to add two workers and correspondingly increase space in the kitchen. Doing so, at moderate cost, provided the following reassurances:

1. Kitchen utilization would decrease from a frenetic 87% to a relatively comfortable but not wasteful 74%-75%.
2. A fivefold increase in customers could be accommodated with unchanged customer waiting time. This reassurance was of particular significance to the restaurateur because new construction of office buildings in the neighborhood is likely to provoke a sharp increase in the number of office workers wishing to order a takeout dinner as they prepare to commute home from work.
3. The time-in-system of a customer (from the customer’s perspective, “How soon after calling my order will it be ready?”) would be less than $\frac{1}{3}$ its previous value.
4. The number of customers served on average increased from 40.55 to 45.00, a 10% increase.

Details of these simulation results appear in Table 2 (Appendix). The box-&-whisker plot (Figure 3), very easily and conveniently available in Simio®, helps a non-technical client, such as the proprietor, readily

understand the merits of the sixteen different scenarios studied relative to the performance metrics (here, customer time in system).

In view of these considerations, the proprietor did adopt this change and successfully realized these benefits, supporting a rapid and profitable increase in business. Financially, the cost of increased capacity was 10,000 Rs. (one time), the cost of the two incremental workers was 8,200 Rs./month, and overall profit increased by Rs. 169,000, a nearly nine-fold return on investment.

CONCLUSIONS AND FURTHER WORK

This simulation project and its analysis provided valuable advice and reassurances to the restaurateur. Deliberate inclusion of a large number of scenarios – many of which were deliberately *leaner* in staffing levels than the current situation – helped convince the restaurateur that the operations were most emphatically *not* overstaffed, and added to the credibility the model and its analysis achieved. Furthermore, the successes of increasing total customers served by 10% while keeping the number of dropped calls negligible and *not* increasing customer waiting time have encouraged the business proprietor to consider potential expansions of the enterprise.

ACKNOWLEDGMENTS

The authors gratefully acknowledge timely support from the company Simio LLC as various technical questions arose during the course of this project. Also, helpful criticisms from three anonymous reviewers have enabled the authors to make meaningful improvements to the organization and presentation of the results of this study.

REFERENCES

- Achaya, K. T. 2012. *The Story of our Food*. India: Universities Press Private Limited.
- Brann, David M. and Beth C. Kulick. 2002. Simulation of Restaurant Operations Using the Restaurant Modeling Studio. In *Proceedings of the 2002 Winter Simulation Conference*, Volume 2, eds. Enver Yücesan, Chun-Hung Chen, Jane L. Snowdon, and John M. Charnes, 1448-1453.
- Chung, Christopher A. 2004. *Simulation Modelling Handbook*. Boca Raton, Louisiana: CRC Press.
- Curin, Sara A., Jeremy S. Vosko, Eric W. Chan, and Omer Tsimhoni. 2005. Reducing Service Time at a Busy Fast Food Restaurant on Campus. In *Proceedings of the 2005 Winter Simulation Conference*, eds. Michael E. Kuhl, Natalie Steiger, Frank Armstrong, and Jeffrey A. Joines, 2628-2635.
- Joines, Jeffrey Allen and Steven Dean Roberts. 2015. *Simulation Modeling with Simio: A Workbook*, 4th edition. North Charleston, South Carolina: CreateSpace Independent Publishing Platform.
- Kelton, W. David, Jeffrey Smith, and David Sturrock. 2013. *Simio and Simulation: Modeling, Analysis, Applications*, 3rd edition. Learning Solutions.
- Law, Averill M. and Michael G. McComas. 1999. Simulation of Manufacturing Systems. In *Proceedings of the 1999 Winter Simulation Conference*, Volume 1, eds. Phillip A. Farrington, Harriet Black Nembhard, David T. Sturrock, and Gerald W. Evans, 56-59.
- Leemis, Lawrence M. 2002. Stat::Fit: Fitting Continuous and Discrete Distributions to Data. In *OR/MS Today* (29,3) [June].
- Sargent, Robert G. 2015. An Introductory Tutorial on Verification and Validation of Simulation Models. In *Proceedings of the 2015 Winter Simulation Conference*, eds. L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 1729-1740.
- Starks, Darrell W., and Todd C. Whyte. 1998. Tutorial: Simulation in the Hospitality Industry. In *Proceedings of the 1998 Winter Simulation Conference*, Volume 1, eds. D. J. Medeiros, Edward F. Watson, John S. Carson, and Mani S. Manivannan, 37-39.
- Thiesing, Renee M. and C. Dennis Pegden. 2015. Introduction to Simio. In *Proceedings of the 2015 Winter Simulation Conference*, eds. L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 4090-4099.
- Williams, Edward J., Justin Clark, Renée M. Amodeo, and Jory D. Bales Jr. 2005. Simulation Improves Staffing Procedure at an Oil Change Center. In *Proceedings of the 19th European Conference on Modelling and Simulation*, eds. Yuri Merkurjev, Richard Zobel, and Eugène Kerckhoffs, 309-314.

AUTHOR BIOGRAPHIES

SAPTHAGIRISHWARAN **THENNAL SIVARAMAKRISHNAN** is pursuing his master's in business analytics, majoring in information management and coordination analytics. He holds a master of business administration from VIT University, Vellore, India. He earned his bachelor's in computer science and engineering from Jerusalem College of Engineering (Anna University), Chennai, India. From 2011 to 2013, he worked as a software engineer for HCL Technologies Limited where his role was to develop and customize software as well as perform base product migration for the U.S.-based pharmaceutical customers. During his tenure with HCL, he worked for the clients CSL Behring LLC and Eisai Incorporated. He works at ILABS, the center for innovation research, University of Michigan – Dearborn as a graduate research assistant where he is conducting data analysis for the College of Arts, Sciences and Letters (CASL) to unleash hidden insights. He is also the webmaster of the Information Technology Management (ITM) club at his university. His email address is stsivara@umich.edu.

SHANMUGASUNDARAM CHANDRASEKARAN pursued his Master's in Business Analytics majoring in Information Management & Coordination Analytics. He holds a Master of Business Administration degree from VIT University, Vellore, India. He earned his

Bachelor's in Electronics and Communication Engineering from K.S. Rangasamy College of Technology (Anna University), Tamilnadu, India. From 2012 to 2013, he held a position as placement coordinator for the ECE Department, K.S. Rangasamy College of Technology where his role was to coordinate group activities related to placement and help students to achieve job placement in highly regarded companies. He works at Petrofac Engineering Service Private Limited as a management intern where he has proposed a few changes in the project change management of the ERP Systems, to improve their processes. In 2013, he published a paper "Sensor Network Based Dyeing Industry Monitoring and Controlling System" in the *International Journal of Teacher Educational Research* (IJTER). In 2011, he won first prize for his paper "Cellphone the Life Saver," presented in Alagappa College of Technology, Tamilnadu, India. His email address is shanmugc@umich.edu.

JENNIFER DHANAPAL is a graduate student in supply chain management enrolled with the College of Business at the University of Michigan-Dearborn. Prior to this she studied business administration as her first graduate degree program at VIT University, India. Her bachelor's degree was in commerce and finance from the University of Madras, India. She has extensive experience in quality assurance and data accuracy, having worked with Frost & Sullivan undertaking consulting and market research. She also has experience working in production and operations with auto manufacturer Royal Enfield. She was instrumental in the automation process of materials costing and helped in improving the efficiency of the procurement system. She is a student member of American Production and Inventory Control Society (APICS) and won third prize in the 2015 Great Lakes supply chain case solving competition organized by APICS. Her email address is jdhanapa@umich.edu.

PAUL AJAYDIVYAN JEYA SEKAR is a graduate student in supply chain management enrolled with the college of business at the University of Michigan-Dearborn. He holds a master of business administration degree from VIT University, Vellore, India. He earned his bachelor's degree in technology from Kalasalingam University, Virudhunagar, India. He published an international paper on nanotoxicity at Christ College, Kerala and conducted international knowledge carnival-gravitas 2013 at VIT University. He also did his internship at Saint Gobain-Weber studying the brand penetration of a new product that has been launched and negotiated with the purchasing decision of top management. He is a student member of the American Production and Inventory Control Society (APICS). His email address is pjeya@umich.edu.

EDWARD WILLIAMS holds bachelor's and master's degrees in mathematics (Michigan State University, 1967; University of Wisconsin, 1968). From 1969 to 1971, he did statistical programming and analysis of biomedical data at Walter Reed Army Hospital, Washington, D.C. He joined Ford Motor Company in 1972, where he worked until retirement in December 2001 as a computer software analyst supporting statistical and simulation software. After retirement from Ford, he joined PMC, Dearborn, Michigan, as a senior simulation analyst. Also, since 1980, he has taught classes at the University of Michigan, including both undergraduate and graduate simulation classes using GPSS/H™, SLAM II™, SIMAN™, ProModel®, SIMUL8®, or Arena®. He is a member of the Institute of Industrial Engineers [IIE], the Society for Computer Simulation International [SCS], and the Michigan Simulation Users Group [MSUG]. He serves on the editorial board of the *International Journal of Industrial Engineering – Applications and Practice*. During the last several years, he has given invited plenary addresses on simulation and statistics at conferences in Monterrey, México; İstanbul, Turkey; Genova, Italy; Rīga, Latvia; and Jyväskylä, Finland. He served as a co-editor of *Proceedings of the International Workshop on Harbour, Maritime and Multimodal Logistics Modelling & Simulation 2003*, a conference held in Rīga, Latvia. Likewise, he served the Summer Computer Simulation Conferences of 2004, 2005, and 2006 as *Proceedings* co-editor. He was the Simulation Applications track coordinator for the 2011 Winter Simulation Conference. A paper he co-authored with three of his simulation students won "best paper in track" award at the Fifth International Conference on Industrial Engineering and Operations Management, held in Dubai, United Arab Emirates, in March 2015. His email addresses are ewilliams@pmcorp.com and williams@umich.edu.

APPENDIX

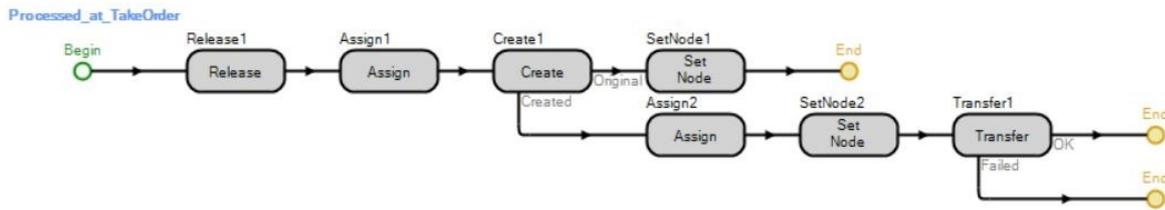
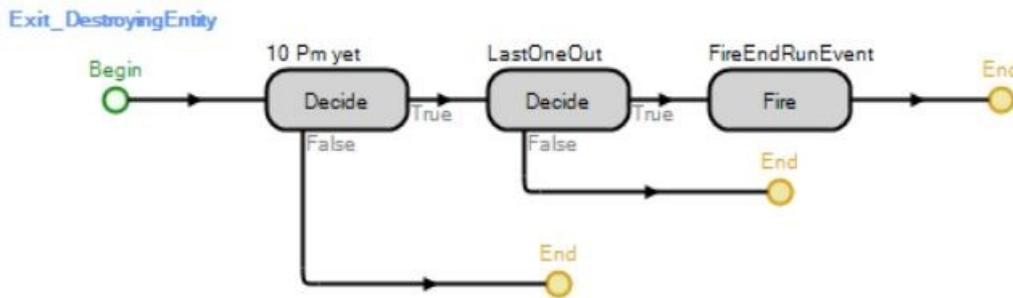


Figure 1. Process Logic Executed when Order Is Taken



CloseTheShop After 10pm, last customer leaving, close shop

LastCustomerLeaving

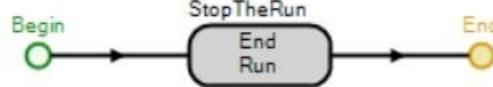


Figure 2. Model Closeout Logic

Table 1. Key Performance Metrics for Sixteen Scenarios

#Lines	#Workers	Lost Calls/hr	Kitchen Util.	Cust TIS (hr)	# Cust in Q	Cust Wait (min)
1	1	0.71	90.03	0.16	0.02	22.68
1	2	0.70	87.58	0.06	0.03	22.49
1	3	0.60	80.12	0.03	0.05	24.44
1	4	0.63	66.81	0.01	0.06	23.71
2	1	0.15	90.13	0.18	0.03	35.00
2	2	0.14	87.32	0.07	0.06	40.00
2	3	0.14	81.73	0.03	0.08	37.01
2	4	0.20	71.24	0.01	0.10	36.63
3	1	0	90.14	0.18	0.03	41.10
3	2	0	87.01	0.07	0.06	43.16
3	3	0.03	81.93	0.03	0.09	40.09
3	4	0.08	73.45	0.02	0.12	42.71
4	1	0	90.14	0.18	0.03	41.10
<i>4</i>	<i>2</i>	0	87.01	0.07	0.06	43.16
4	3	0	81.53	0.03	0.09	40.00
4	4	0.03	74.49	0.02	0.12	43.12

Red italics: Original situation. Blue boldface: Recommended and adopted situation.

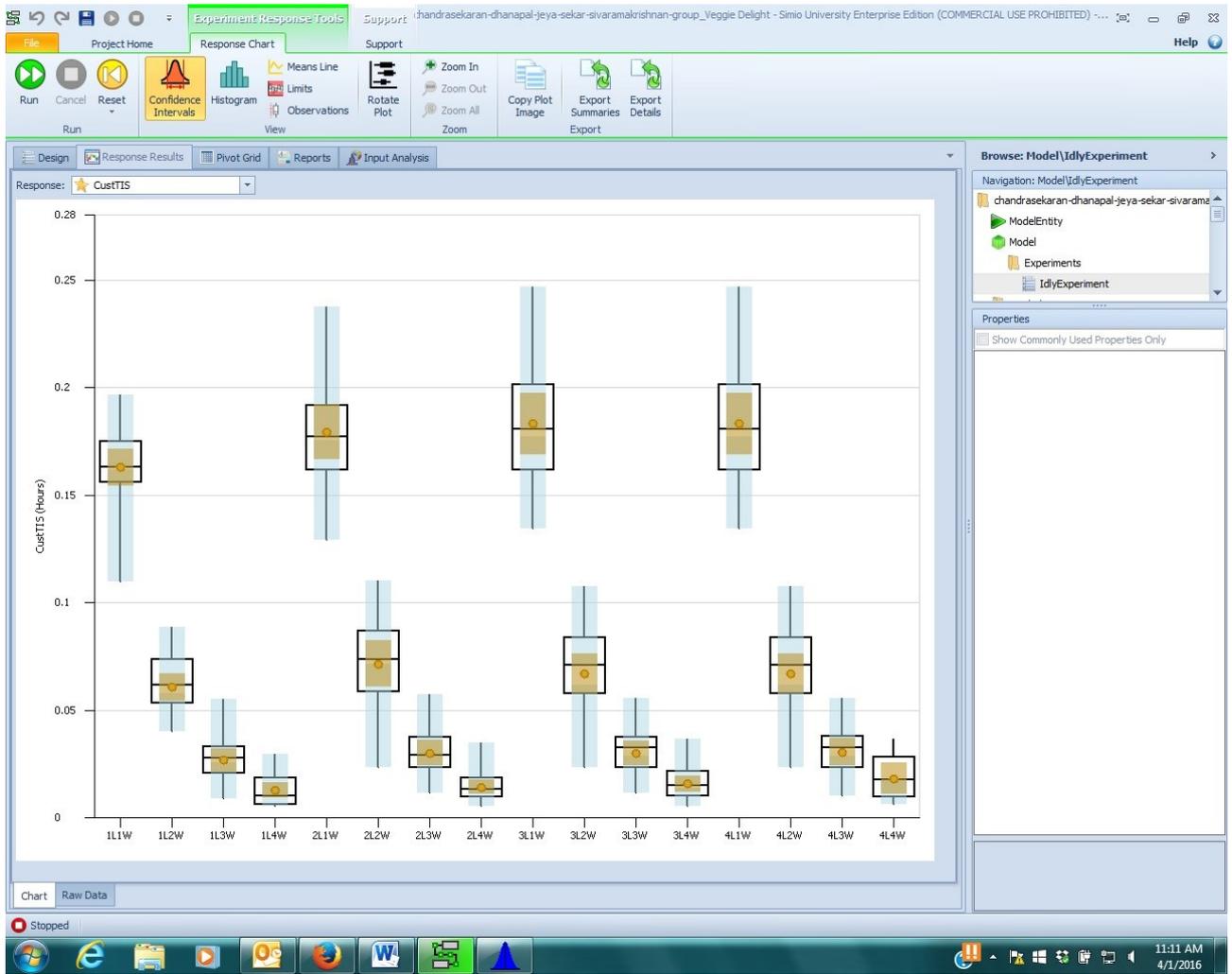


Figure 3. Comparative Box-&-Whisker Plots for Customer Time-In-System Under Sixteen Scenarios

MAKING OF CREDIBLE PERMEABILITY MAPS FOR LAYERS OF HYDROGEOLOGICAL MODEL OF LATVIA

Aivars Spalvins, Inta Lace, Kaspars Krauklis
Environment Modelling Centre
Riga Technical University
Daugavgrivas str. 2, Riga, LV-1007, Latvia
E-mail: emc@cs.rtu.lv

KEYWORDS

Hydrogeological model, numerical interpolation, permeability of geological layers, pumping tests for wells, transmissivity of aquifers.

ABSTRACT

In 2010–2012, the hydrogeological model (HM) of Latvia (LAMO) was developed by the scientists of Riga Technical University (RTU). LAMO comprises geological and hydrogeological data provided by the Latvian Environment, Geology and Meteorology Centre (LEGMC) for the active groundwater zone of Latvia. In 2013–2015, LAMO was notably upgraded. The density of hydrographical network (rivers, lakes) was increased, cuttings of river valleys into primary geological layers were done, plane approximation step was decreased, hydraulic conductivity distributions of layers were refined by creating more reliable permeability maps. In the paper, methods of obtaining these maps are described.

INTRODUCTION

The countries of the European Union (EU) are developing the HM from which information is applied for the water resources management that must implement the EU aims defined in the Water Framework Directive (Water Framework Directive 2000). In Latvia, the LEGMC specialists are preparing and updating the water resources management plans for the country.

In 2010–2012, the HM LAMO was established by the scientists of RTU. LAMO simulates the steady state average hydrogeological simulation of Latvia. The licensed program Groundwater Vistas (GV) is used for running LAMO (Environmental Simulations, Inc. 2011).

In 2013–2015, LAMO was upgraded (Spalvins et al. 2015a). Due to these upgrades, four successive versions of LAMO can be marked (see Table 1).

LAMO comprises the active groundwater zone of Latvia that provides drinking water. In Figure 1, the location of LAMO is shown.



Figure 1: Location of LAMO.

The land territory of Latvia and the area of the Gulf of Riga constitute the HM active area (Figure 2). The passive area represents border territories of the neighbouring countries. The active and passive areas are separated by the 4km wide border zone where boundary conditions for the active area are fixed.

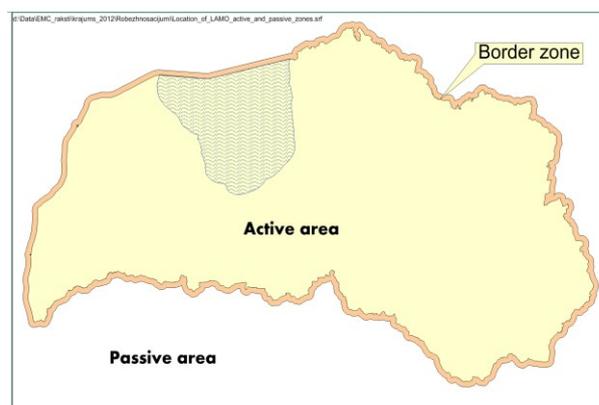


Figure 2: Locations of LAMO active and passive areas

Table 1: Versions of LAMO

Name of version	Year of disposal	Approximation grid			Rivers in model			Lakes number
		Plane step of [metre]	Number of grid planes	Number of cells [$\times 10^6$]	Number	Valleys incised	Flow data used	
LAMO1	2012	500	25	14.25	199	no	no	67
LAMO2	2013	500	27	15.43	199	yes	no	67
LAMO3	2014	500	27	15.43	469	yes	no	127
LAMO4	2015	250	27	61.56	469	yes	yes	127

The LAMO4 version simulates 27 geological layers (see Table 2). It is shown in Figures 3 and 4 that most of primary layers are outcropping. After emerging at the sub quaternary surface, such layers have zero thickness ($m = 0$).

Table 2. Vertical schematization and parameters of layers for calibrated LAMO4

No of HM layer	*	Name of layer	HM layer code	Area, [thous. km ²]	m_{mean} , [meter]	k_{mean} [meter/day]
1		Relief	relh	71.29	0.02	10.0
2		Aeration zone	aer	71.29	0.02	3.1×10^{-6}
3		Unconfined Quaternary	Q2	71.29	5.77	11.2
4		Upper moraine	gQ2z	71.29	22.20	1.4×10^{-3}
5		Confined Quaternary	Q1#	7.4	6.13	7.0
6		Lower moraine	gQ1#z	9.7	9.3	2.8×10^{-4}
7		Ketleru	D3ktl#	5.32	61.46	4.2
8		Ketleru	D3ktlz	5.79	10.52	2.8×10^{-4}
9		Zagares	D3zg#	7.43	42.65	7.0
10		Akmenes	D3akz	7.95	11.05	2.8×10^{-5}
11		Kursas	D3krs#	9.34	22.34	6.3
12		Elejas	D3el#z	9.24	27.58	2.8×10^{-5}
13		Daugavas	D3dg#	32.14	30.37	9.4
14		Salaspils	D3slp#z	35.78	12.67	8.4×10^{-4}
15		Plavīnu	D3pl	43.80	22.76	8.6
16		Amatas	D3am#z	45.14	8.97	1.4×10^{-4}
17		Amatas	D3am	46.21	21.91	6.4
18		Upper Gauja	D3gj2z	48.80	11.62	2.8×10^{-4}
19		Upper Gauja	D3gj2	50.92	26.34	6.2
20		Lower Gauja	D3gj1z	53.11	13.17	2.8×10^{-4}
21		Lower Gauja	D3gj1	56.13	31.55	5.4
22		Burtņieku	D2brtz	58.09	15.41	5.6×10^{-4}
23		Burtņieku	D2brt	68.74	45.02	4.2
24		Arikula	D2arz	68.74	15.02	4.2×10^{-4}
25		Arikula	D2ar	68.74	40.03	3.2
26		Narva	D2nr#z	71.29	116.67	2.8×10^{-5}
27		Pemava	D2pr	71.29	25.00	10.0

*  - aquitarid

m_{mean} and k_{mean} – the mean thickness and permeability

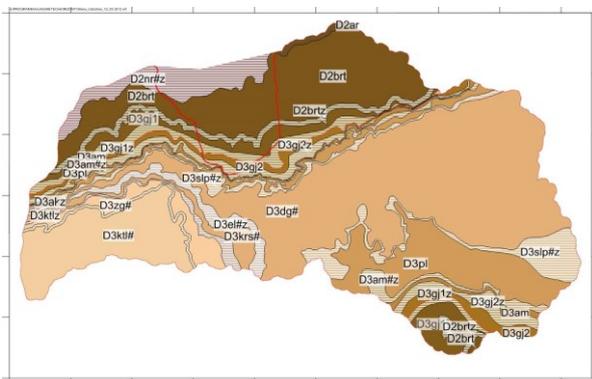


Figure 3: Boundaries of primary geological strata

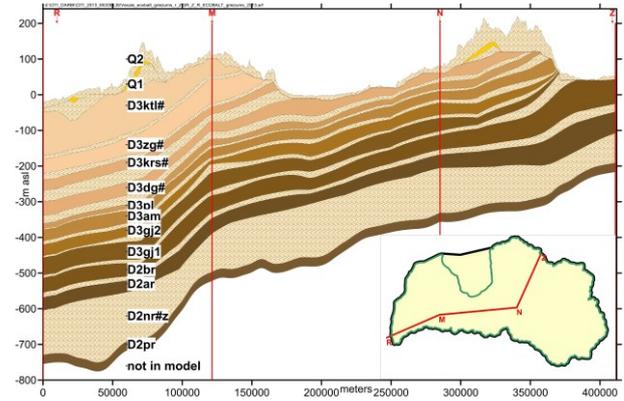


Figure 4: Geological cross section

The layers Nos 1, 2, 3, 4, 26, 27 do not have the $m = 0$ areas, because they exist everywhere in the HM area of the active part. The size of the $m = 0$ areas of the LAMO layers can be computed by using the data of Table 2. as the difference between the area of the HM active part (71.29 thous.km²) and the $m > 0$ area of the layer. The $m = 0$ areas caused problems when the permeability maps (k -maps) for layers of LAMO were obtained (Spalvins et al. 2015b).

In the present paper, methods are described that were used for making more realistic k -maps of LAMO4.

MATHEMATICAL FORMULATIONS

To understand the problems of creating the k -maps for the $m > 0$ and $m = 0$ areas of HM layers, the basic mathematics of HM must be considered

By using the 3D - finite difference approximation, the xyz -grid of HM is built. It consists of $(h \times h \times m)$ -sized blocks (h is the block plane step, m is the variable thickness of a geological layer). For LAMO4, $h = 250$ meters.

LAMO provides the 3D-distribution of piezometric head vector φ as the numerical solution of the boundary field problem which is approximated in the nodes of the HM xyz -grid by the following algebraic expression (Spalvins et al. 2015a):

$$A\varphi = \beta - G\psi, \quad A = A_{xy} + A_z, \quad (1)$$

where A is the hydraulic conductivity matrix of the geological environment which is presented by the xy -layer system containing horizontal (A_{xy} -transmissivity T) and vertical (A_z - vertical hydraulic conductivity) elements of the HM grid; ψ and β are the boundary head and flow vectors, respectively; G is the diagonal matrix (part of A) assembled by elements linking the nodes where φ must be found with the locations where ψ is given.

The ψ -conditions for LAMO are fixed on all outer boundary surfaces of the HM active volume (top, bottom and the vertical surface of the shell in the

boundary zone). The ψ -conditions also include conditions for rivers and lakes. In LAMO, the β -conditions are used only for exploitation wells. The elements a_{xy} , a_z of A_{xy} , A_z (or g_{xy} , g_z of G) for the block ($h \times h \times m_i$) are computed as follows:

$$\begin{aligned} a_{xyi} &= k_i m_i = T_i, & a_{zi} &= h^2 k_i / m_i, \\ m_i &= z_{i-1} - z_i \geq 0, & i &= 1, 2, \dots, u, \end{aligned} \quad (2)$$

where z_{i-1} , z_i are elevations, accordingly, of the top and bottom surfaces of the i -th geological layer; z_0 represents the ground surface elevation ψ_{rel} -map; m_i , k_i are elements of digital m_i , k_i -maps of thickness and permeability of the i -th layer, accordingly; for LAMO4, $u=28$. The set of z -maps describes the full geometry (stratification) of LAMO.

For the block (cell) of the xyz -grid, the surfaces z_{i-1} and z_i represent its top and bottom, correspondingly. The centres of the cells are located on the surface $z_{i-0.5}$:

$$z_{i-0.5} = 0.5(z_{i-1} + z_i), \quad i=1, 2, \dots, u. \quad (3)$$

The vertical link $a_{zi,i+1}$ that joins the centers of the i -th and the underlying $(i+1)$ -th cell is computed, as the harmonic mean of the conductances a_{zi} and a_{zi+1} :

$$a_{zi,i+1} = 2 a_{zi} a_{zi+1} / (a_{zi} + a_{zi+1}), \quad (4)$$

where a_{zi+1} is the vertical hydraulic conductivity of the $(i+1)$ -th cell. It follows from (2) that for the $m=0$ areas of layers: $a_{xy} = k \times 0 = 0$ and $a_z = k/0 = \infty$.

In the GV system, LAMO is supported by the MODFLOW program (Harbaugh 2005), where the matrix A must be simulated accurately also for the $m=0$ areas. In order to match this rule and to avoid the "division by zero" in the a_z calculation, $m=0$ must be replaced by a small $\varepsilon > 0$. For LAMO, $\varepsilon = 0.02$ meter.

It follows from (2), the a_{xy} and a_z -maps are computed by the GV system where the k and z -maps serve as the initial data. The z -maps simulate the geological stratigraphy that cannot be changed easily. In fact, the a_{xy} and a_z -maps can be controlled only by altering the k -maps.

For the $m=0$ areas, problems arise if the k -maps for aquifers are obtained by using the formula:

$$k = T / m, \quad (5)$$

where the m -map is the divider. For the $m=0$ areas of aquitards, obstacles of obtaining their k -maps also must be removed.

PERMEABILITY MAPS FOR AQUITARDS

The basic indication of an aquitard which impedes flow of groundwater is its small value of k (see Table 2). Due to this fact, the transmissivity $a_{xy} = T$ of aquitards has no real influence on the solution φ of (1). Due to smallness

of a_{xy} , no boundary conditions are used for aquitards. As it follows from (4), the vertical link a_{zi} of an aquitard takes over the link a_{zi+1} of an aquifer, because $a_{zi+1} \gg a_{zi}$ and $a_{zi,i+1} \sim a_{zi}$. For this reason, the aquitards control vertical groundwater flows between the aquifers.

The k -maps of the aquitards of LAMO were obtained by using general knowledge about aquitards of Latvia and by adjusting the maps in the course of calibration.

Special correction of the k -maps was done for the $m=0$ areas of the ten aquitards Nos 6, 8, 10, 12, 14, 16, 18, 20, 22, 24. At the northern part of Latvia, their $m=0$ areas are overlapping with those of the eleven aquifers Nos 5, 7, ..., 25. (see Table 2, Figures 3 and 4). For the 21 cells of the size ($h \times h \times \varepsilon$), the series conductance a_{zn} of the volume ($h \times h \times 21\varepsilon$) can be computed by using the expression for the series connection of a_{zi} :

$$a_{zn} = 1 / \sum_{i=1}^n (1/a_{zi}) = (h^2 / \varepsilon) / \sum_{i=1}^n (1/k_i)_{mean} = 32, \quad (6)$$

where $k_i = k_{mean}$ are the data in Table 2; $h = 250$; $n = 21$. It is obvious that the value $a_{zn} = 32$ cannot be accepted, because it is very far from $a_z = \infty$. To mend a_{zn} at the $m=0$ areas, the k_{mean} was increased 100 and 10 times, for the aquitards Nos 8, 10 and Nos 12-24, accordingly. After the correction, the value of k_{zn} increased from the unacceptable 32 to 934.

Within the $m > 0$ area of an aquitard, the $m=0$ track can appear if the aquitard is cut through by a river valley. These tracks are treated as the ordinary $m=0$ areas.

The choice of $\varepsilon = 0.02$ meter was determined also by the tolerable error of stratification in the $m=0$ areas for the northern part of Latvia. There the total thickness of the $m=0$ volume is 21ε . If $\varepsilon = 0.02$ then there the HM geometry distortion 1.02 meter is acceptable.

PERMEABILITY MAPS FOR AQUIFERS

The transmissivity a_{xy} for aquifers is very important, because it controls the lateral groundwater motion there. Because a_z of aquifers have large values, they have small influence on vertical links of (4), hence in LAMO they join aquifers with aquitards.

The permeability of aquifers can be found in a variety of ways: field tests, laboratory tests, methods based on grain size distributions (Domenico and Schwartz 1998). Inverse problem solving methods can also be used (Chin 2014). However, the field tests where one well is pumped are commonly applied. They permit the testing of large volumes of rock. They have provided rather reliable data for finding permeability of aquifers for the LAMO3 and LAMO4 versions

It was shown in (Spalvins et al. 2015b), how the variable k -maps were obtained for the LAMO3 version by using the pumping data of wells. New methods that have been used for creating more reliable k -maps for the LAMO4 version are described in this section.

Appliance of pumping data of wells

The pumping test of a single well in a confined aquifer uses the discharge rate Q . The drawdown S of the groundwater head is observed which value is given by the expression (Bindeman and Jazvin 1982):

$$S = \frac{Q}{2\pi T} (\ln(R/r) + \xi + \gamma), \quad T = km, \quad (7)$$

where R and r are radiuses, accordingly, of the well depression cone and the screen; ξ and γ are dimensionless hydraulic resistances that account for the partial penetrating factor of a well and for the quality of the well screen, respectively. For a new well, $\gamma = 0$. For old wells, the screen resistance γ increases; its value is unknown and, for this reason, only pumping data of the new wells can provide credible results. Thus, $\gamma = 0$ should be used in (7).

From (7), the following expression can be obtained:

$$T = \frac{q}{2\pi} (\ln(R/r) + \xi), \quad q = Q/S, \quad (8)$$

where q is the specific capacity of a well.

If q and T have the dimensions, liter/(sec.meter), and (meter)²/day, respectively, then

$$T = 13.75q(\ln(R/r) + \xi). \quad (9)$$

It was shown in (Spalvins et al. 2015b) that for the leaky confined primary aquifers of LAMO, $\ln(R/r) \sim 10.0$. If $\xi=0$ then (9) is roughly approximated by the expression:

$$T_{\min} = 137q. \quad (10)$$

In (Verigin 1962), the empirical formula is given for obtaining ξ :

$$\xi = (1/a - 1)(\ln 1.47ab - 2.65a), \quad a = l/m, \quad b = m/r, \quad (11)$$

where m is the thickness of an aquifer, and l and r are, accordingly, the length and radius of the well screen. The formula can be used if $m/r > 100$, $l/m \geq 0.1$.

The resistance ξ can be applied to refine the transmissivity T , as follows:

$$T = \nu T_{\min}, \quad \nu = 1 + \xi/10.0, \quad k_{\text{cor}} = \nu k, \quad (12)$$

where k_{cor} – the corrected value of k .

For LAMO, the typical values of l/m and m/r are within the limits: $0.5 > l/m > 0.2$; $500 > m/r > 100$. Then, as follows from (Spalvins et al. 2015b), the correction factor ν may be within the limits: $2.8 > \nu > 1.3$.

Presently, ξ is not accounted for. However, (12) shows that a modeler can use $T > T_{\min}$, if necessary. It was done for the LAMO4 version (Spalvins et al. 2015a). In Table 2, k_{mean} have larger values than the ones in Table 4 where $\nu = 1.0$.

Obtaining of permeability maps

The k -maps for aquifers can be obtained by using the formula (5) where the transmissivity T is derived from the well pumping data; m is the aquifer thickness which is used in (2) by the GV system.

By using the Excel program (Walkenback 2007), the set of the specific capacity q , must be extracted from the well pumping data. As a rule, the q -set contains very low and also very high improbable values. In order to normalize the set, minimal and maximal values of q are fixed (for LAMO3, $q_{\min} = 0.3$ and $q_{\max} = 5$). The q -set contains n pointwise data. For LAMO, $n > 1000$ for practically all aquifers. Due to the large n , the fast gridding method of “inverse distance to power” is applied by the SURFER program (Golden software, Inc 2012) This method computes the interpolated value σ_o at the grid nodes by using the available pointwise data $\sigma_i = q_i$, $i = 1, \dots, n$, as follows (Franke 1982):

$$\sigma_o = \left(\sum_{i=1}^n \sigma_i \tau_i \right) / \sum_{i=1}^n \tau_i, \quad \tau_i = (1/d_{oi})^p,$$

$$d_{oi} = \sqrt{(x_o - x_i)^2 + (y_o - y_i)^2}, \quad (13)$$

where τ_i – the weight of σ_i ; d_i – the distance between the grid node o and the σ_i point; p – the weighting power; $x_o, y_o; x_i, y_i$ are coordinates, respectively, of the o -th grid node and the i -th point. The value $p = 2$ was used to prepare the σ -grid for LAMO3 and LAMO4.

The interpolation result of (13) is rather rough and, to smooth it, the moving digital “inverse distance” low-pass filter of the size 11×11 was used (Spalvins et al. 2015b), (Ditas 2000):

$$\sigma_{oo} = \left(\sum_{i,j} \sigma_{ij} \tau_{ij} \right) / \sum_{i,j} \tau_{ij}, \quad \tau_{ij} = (1/D_{ij})^p,$$

$$D_{ij} = \sqrt{i^2 + j^2}, \quad (14)$$

where τ_{ij} – the weight of σ_{ij} ; p – the power ($p = 0.5$ was applied); i and j were the grid row and column local indices for the neighboring nodes with respect to the central node oo of the filter; D_{ij} – the distance between the nodes oo and ij .

Smoothing of $\sigma = q_{ij}$ by the filter (14) is moderate. To preserve the data provided by wells, only one filtering pass was done.

The “inverse distance” interpolation and filtering do not account for discontinuity of aquifers that include the $m = 0$ areas. Therefore, for all nodes of the LAMO grid, values of q_{ij} are computed. Because the formula (5) are used, only at the $m > 0$ area, reasonable k values may appear.

Very large k values appear within the $m = 0$ areas (there $m = 0.02$ meter for LAMO), at a vicinity of borderlines within the $m > 0$ areas where $m \rightarrow 0$. At locations of river valleys, the values of k jumpwise enlarge, due to the decrease of m at the valley places.

For LAMO3, the extreme k values at the $m = 0$ and $m \rightarrow 0$ areas were replaced by the maximal k value that was found within the $m > 0$ zone of the k -map (Spalvins et al. 2015b). No satisfactory method was found to eliminate the jumpwise changes of k at the locations of river valleys.

For LAMO4, the both above-mentioned drawbacks were eliminated. Initial data for q were checked by the computer based tools.

Checking of well pumping data for LAMO4

Pumping data of wells were provided by LEGMC. These data were never checked before. In the case of LAMO3, only rough testing and sorting of them were done (elimination of obviously wrong data, appliance of data bounded within the $5 > q > 0.3$ interval). For the case of LAMO4, more careful checking of data was done. Its results are presented in Table 3, where four stages of the initial data treatment are shown (deposited, sorted, bounded, and surviving) for the ten primary aquifers of LAMO. Table 3 gives the number of wells in each stage and the mean value q_{mean} of the specific capacity of wells that are present at the stage.

The value of q_{mean} is the arithmetic mean:

$$q_{mean} = \left(\sum_{i=1}^N q_i \right) / N, \quad (15)$$

where q_i – the specific capacity of i -th well; N – the number of wells.

As it follows from Table 3, a rather large number of wells were not allowed to take part at the second stage “selected”. The eliminated wells were with obviously wrong data and the ones which screens were not located entirely within the aquifer under the pumping test. If a screen is located in two or more aquifers then the well cannot be used for finding its q or the real piezometric head of the aquifer (Tremblay et al. 2015).

For the aquifers D3gj1 and D3gj2, due to this feature, a considerable part of their wells were not accepted for the second stage (see Table 3).

During the third stage “bounding”, the wells are eliminated which q does not belong to the interval $4 > q > 0.2$. The value of q_{mean} increases for all aquifers, because the number of wells ($q < 0.2$) is much larger than the ones ($q > 4$).

To perform the fourth stage “surviving”, two sequential steps are carried out:

1-st step: within a circle of the radius R_1 , only the one well remains which q is the largest;

2-nd step: within a circle of the radius R_2 , the wells remain which hold the condition $(1 + \Delta) > q_{mean} > (1 - \Delta)$ where q_{mean} is computed within the circle and Δ is the deviation from the value of q_{mean} .

During the first step, the wells with contradictory data were eliminated, and the wells with the locally larger q were saved. The second stage is more conservative, because more than one well may be saved within the circle of the radius R_2 .

Table 3: Summary of well data treatment

Aquifer code	Number of wells				q_{mean}		
	deposited	selected	bounded	surviving	selected	bounded	surviving
D3ktl#	288	156	114	46	0.72	0.79	0.88
D3zg#	872	681	533	143	0.80	0.87	1.08
D3krs#	712	524	426	118	0.84	0.86	1.11
D3dg#	2284	959	819	256	1.17	1.15	1.74
D3pl	2874	1295	1073	374	1.08	1.05	1.46
D3am	778	526	420	190	0.64	0.71	0.80
D3gj2	5241	1229	1096	324	0.77	0.84	1.05
D3gj1	5346	1579	1378	425	0.82	0.88	1.18
D2brt	1867	1332	1020	367	0.71	0.80	0.99
D2ar	1740	1188	974	314	0.64	0.71	0.88

For LAMO4, the following search parameters were used: $R_1 = 2000$ meters, $R_2 = 4000$ meters, $\Delta = 0.3$. As it follows from Table 3, in the stage “surviving”, the number of wells is considerably reduced. The value of q_{mean} increased, because for the both steps of the fourth stage, the wells with the locally larger q were saved.

Correction of permeability maps for aquifers

The primary aquifers of LAMO have the $m = 0$ areas. Some of them are cut by river valleys

It was noted above that incisions of river valleys into primary layers caused jumpwise increases of k for the LAMO3 k -maps. This drawback was completely eliminated, because the m_0 -maps without the incisions were used for the LAMO4 case. Such m_0 -maps have been applied by the LAMO1 version, and they are used even nowadays as the starting position for all necessary changes in the HM geometry (set of z -maps).

Appliance of the m_0 -maps is founded on the assumption, that a river valley does not change k .

To suppress the extreme k values, for the $m \rightarrow 0$ zone, the following correction matrix C was used:

$$1 > C = m_0 / (0.75 m_{mean}) \geq 0, \quad (16)$$

where the factor 0.75 was chosen empirically; within the $m = 0$ and $m > 0$ areas, $C=0$ and 1, respectively. The corrected q_{cor} , k_{cor} and T are obtained, as follows:

$$q_{cor} = C q, \quad k_{cor} = 137 q_{cor} / m_0, \quad T = k_{cor} m. \quad (17)$$

In (17), the real m -map is used for obtaining of T and, at locations of river valleys, the values of T take jumpwise decreases, as it must be.

For the $m > 0$ area of an aquifer, the mean arithmetical value k_{mean} of k_{cor} must be found. Within the $m = 0$ area, $k_{cor} = 0$ must be replaced with k_{mean} . The replacement secures the space continuity of HM in the z -direction. In

nature, the continuity of the geological environment is secured by its zero thickness $m = 0$.

In LAMO, the matrix K_{norm} that results from (17), is used as the product:

$$K_{\text{cor}} = K_{\text{norm}} / k_{\text{mean}}, \quad (18)$$

where $k_i = 1.0$, in the $m = 0$ areas of K_{norm} .

In order to decrease a_{xy} in the $m = 0$ areas, (ideally. there $a_{xy} = 0$), K_{cor} elements there are multiplied by 0.1.

Summary on the k-maps

In Table 4, the summary on the k -maps for primary aquifers of the LAMO2, LAMO3 and LAMO4 versions is given. For each HM version, k_{mean} and $k_{\text{max}}/k_{\text{mean}}$ are presented. For the LAMO2 version, $k_{\text{max}}/k_{\text{mean}} = 1.0$, because constant values of k were used for all aquifers. For the LAMO3 and LAMO4 versions, the ratio $k_{\text{max}}/k_{\text{mean}}$ is variable. For LAMO4, the ratio $k_{\text{max}}/k_{\text{mean}}$ is larger than for the LAMO3 version, because the values $q_{\text{min}} = 0.2$ and 0.3 were used for bounding of the initial data of LAMO3 and LAMO4, correspondingly.

Table 4: Summary on LAMO2, LAMO3 and LAMO4 k-maps of the primary aquifers

Aquifer code	LAMO2		LAMO3		LAMO4	
	k_{mean} meter/day	$k_{\text{max}}/k_{\text{mean}}$	k_{mean} meter/day	$k_{\text{max}}/k_{\text{mean}}$	k_{mean} meter/day	$k_{\text{max}}/k_{\text{mean}}$
D3ktl#	3.0	1.0	2.12	9.0	1.77	12.10
D3zg#	3.0	1.0	3.64	5.33	3.38	15.75
D3krs#	2.0	1.0	5.95	4.35	6.33	9.89
D3dg#	10.0	1.0	5.58	14.38	9.40	16.06
D3pl	10.0	1.0	6.11	8.51	8.60	19.65
D3am	10.0	1.0	4.69	5.67	4.64	11.25
D3gj2	10.0	1.0	5.58	4.55	5.11	20.05
D3gj1	14.0	1.0	5.24	6.25	4.84	16.00
D2brt	5.0	1.0	1.91	5.83	3.19	13.75
D2ar	5.0	1.0	2.13	6.15	2.91	17.69

For the LAMO3 and LAMO4 versions, the use of more realistic k-maps for primary aquifers caused changes of their groundwater flow balances (Spalvins et al. 2015a).

CONTROL OF PERMEABILITY MAPS

In MODFLOW and LAMO, the k and m -maps, from the viewpoint of mathematics, are diagonal matrices which elements exist only in nodes of the xyz – grid of HM. For the matrix A of (1), the elements a_{ij} that join the neighbouring nodes with the indices i and j , are computed by MODFLOW as the harmonic mean of a_{xy} , a_z of (2), accordingly, for the lateral and vertical links. No connections exist between the grid nodes that are not neighbours and these links have the zero value.

Due to this short-range feature of the geological environment, the matrix A is sparse, because it contains only $7N \ll N^2$ nonzero elements (N is the number of nodes; N^2 is the number of elements of A). The matrix A is symmetric, because its elements $a_{ij} = a_{ji}$ (Strang 1976).

In LAMO, the final k -map is represented by the diagonal matrix K that is the product of the six factors:

$$K = K_1 \times K_2 \times \dots \times K_6, \quad (19)$$

Table 5 provides the summary on appliance of these six factors K_i for aquitards and aquifers. The factor k_{mean} and the identity matrix I are scalars. The matrix I acts like multiplication by 1; the symbol “+” indicates appliance of the corresponding factor.

Table 5: Factors for controlling the k-maps of LAMO

Factor		Aquitards		Aquifers	
code	action	others	aer	others	Q2
K_1	core matrix	I	1 and 0.05	K_{norm}	K_{norm}
K_2	k_{mean}	+	+	+	+
K_3	change of k	+	+	+	+
K_4	alter k for $m = 0$	1 and 10^2	I	1 and 0.1	I
K_5	alter k for shell	1 and 10^3	1 and 10^3	I	I
K_6	change of m	I	+	I	+

The role of the four factors $K_1 - K_4$ is similar for the all HM layers:

- K_1 is the core matrix; for aquitards and aquifers, accordingly, $K_1 = I$ and K_{norm} ; for the aer zone, the values 0.05 are applied for areas of swamps and locations of lakes and rivers;
- $K_2 = k_{\text{mean}}$ for all layers;
- K_3 is the matrix that is variable during the HM calibration; K_3 is created by using the original GDI (Geological Data Interpolation) program that applies lines as data carriers (Spalvins et al. 2013);
- K_4 accounts for the necessary changes of k for the $m = 0$ areas of layers;
- K_5 is used only for the $m > 0$ areas of the border zone of aquitards where the factor 1000 turns the HM shell into the interpolation tool of the ψ -conditions (Spalvins et al. 2013);

The treatment of the layers Nos 2 and 3 (aer and Q2) after the HM calibration is special, because their thicknesses may be changed, if necessary. During the calibration, $m_{\text{aer}} = 0.02$ meter (see Table 2). The layer aer is the formal aquitard that controls the infiltration flow on the HM top. It is shown in (Spalvins et al. 2013) how the thicknesses m_{aer} and m_{Q2} can be changed, if the appliance of the real m_{aer} is needed. Then the factor K_6 is used.

The aquifers Nos 1 and 27 carry the ψ -conditions and no special data are needed to control them. For this reason, all their six factors K_i are I .

During the HM calibration, the factors K_2 and K_3 must be adjusted..

CONCLUSION

For the present LAMO4 version, the permeability maps were considerably upgraded both for aquifers and aquitards. The more realistic maps of aquifers were created by accounting for pumping data of exploitation wells. The data were checked by computer – based inventory tools. Some drawbacks of the maps that were used in the previous LAMO versions were eliminated. The methods that were used for creating of the LAMO4 permeability maps would be applied by modelers dealing with large regional hydrogeological models.

ACKNOWLEDGEMENTS

In 2010-2012, the hydrogeological model of Latvia LAMO was developed within the framework of the Riga Technical University project that was co-financed by the European Regional Development Fund. The current upgrades of LAMO are supported by the Latvian State Research program “ENVIDEnT” .

REFERENCES

- Bindeman N., and I. Jazvin. 1982. “Evaluation of Groundwater Resources”, Moscow: Nedra, 1982 (in Russian).
- Chin Y.C. 2014. “Application of differential evolutionary optimization methodology for parameter structure identification in groundwater modelling”. *Hydrogeology Journal*, Volume 22, Number 8, December 2014, pp. 1731-1748.
- Ditas I., 2000. “Digital Image Processing Algorithms and Applications”, New York: John Wiley and Sons, 2000.
- Domenico P. A. and F.W. Schwartz. 1998. “Physical and Chemical Hydrogeology”, John Wiley and sons, Inc. – 2 ed. 1998, New York, p. 506.
- Environmental Simulations, Inc. 2011. “Groundwater Vistas. Version 6, Guide to using,” 2011, [Online]. Available: http://www.groundwatersoftware.com/groundwater_vistas.htm
- Franke R. 1982. “Scattered Data Interpolation: Test of Some Methods,” *Mathematics of Computations*, vol.33, pp. 181-200, 1982
- Golden Software, Inc. 2012. “SURFER-11 for Windows, Users manual, Guide to Using,” 2012.
- Harbaugh A.W. 2005. “MODFLOW-2005, U.S. Geological Survey Modular Ground-Water Model: the ground-water flow process”, chap 16, book 6, US Geological Survey Techniques and Methods 6-A16, USGS, Reston, VA.
- Spalvins A., J. Šlangens, I. Lace, O. Aleksans, K. Krauklis , V. Skibelis, I. Eglite. 2015a. “The novel updates of Hydrogeological of Latvia.” *Scientific Journal of Riga Technical University, Boundary Field Problems and Computer Simulation*, vol.54, 23-34, ISSN 2255-9124

http://www.emc.rtu.lv/issues/2015/04_VMC_DITF_54_2015_Spalvins.pdf

- Spalvins A., O. Aleksans, I. Lace. 2015b. “Improving of transmissivity maps for hydrogeological model of Latvia”, *15th international multidisciplinary scientific geoconference (SGEM 2015), June 18–24, 2015*, Albena, Bulgaria, Conference Proceedings, 2015, vol. 1, pp. 667–684

http://www.emc.rtu.lv/issues/2015/BOOK_2_VOLUME_1_IMPROVING_OF_TRANSMISSIVITY.pdf

- Spalvins, A., J. Slangens, I. Lace, K. Krauklis, and O. Aleksans, 2013. “Efficient Methods Used to Create Hydrogeological Model of Latvia,” *International Review on Modelling and Simulations (I.R.E.M.O.S)*, vol. 6, Nr. 5, Okt. 2013., pp. 1718-1726, ISSN 1974-9821, Extracted by ICOMOS 2013
- Strang, G. 1976. *Linear algebra and its applications* / Academic Press, New York, p. 373 INC.
- Tremblay Y., J. M. Lemieux, R. Fortler, J. Molson, R. Therrien, P. Therrien, G. Comeau, M. C. Talbot Poulin. 2015. “Semi- automated filtering of data outliers to improve spatial analysis of piezometric data,” *Hydrogeology Journal*, vol. 23, no. 5, Aug.2015, Springer –Verlag Berlin Heidelberg, pp. 851–868.
- Verigin N. 1962. “Methods Used for Finding Permeability of Geological Strata”, Moscow: Gosstroizdat, 1962 (in Russian).
- Walkenback J. 2007. “Excel-7 Bible”, Indianapolis, Wiley Publishing, Inc., 2007
- Water Framework Directive. 2000. (2000/60/EC of the European Parliament and of the Council). *Official Journal of the European Communities*, L327, 22.12.2000.

AUTHOR BIOGRAPHIES



Aivars Spalvins was born in Latvia. In 1963, he graduated from the Riga Polytechnical Institute (since 1990 – Riga Technical University) as a Computer Engineer. A. Spalvins is the Head of the Environment Modelling Centre of RTU. His research interests include computer modelling of groundwater flows and migration of contaminants.
E-mail: aivars.spalvins@rtu.lv



Kaspars Krauklis received the Master’s degree in Computer Systems from the Riga Technical University in 2007 and the Certificate in Teaching of Engineering Sciences from the Institute of Humanities of RTU in 2005. He is a researcher at the Environment Modelling Centre of RTU.
E-mail: kasparskrauklis@gmail.com



Inta Lace was born in Latvia. In 1971, she graduated from the Riga Polytechnical Institute (since 1990 – Riga Technical University) as a Computer Engineer. In 1995, I. Lace received the Master’s degree in Applied Computer Science. Since 1991, she is a researcher at the Environment Modelling Centre of the Faculty of Computer Science and Information Technology, RTU.
E-mail: intalace@yahoo.com

CONCEPT HIERARCHIES FOR SENSOR DATA FUSION IN THE COGNITIVE IoT

Franco Cicirelli and Giandomenico Spezzano
Institute for High Performance Computing and Networking (ICAR)
CNR - National Research Council of Italy
87036 Rende(CS) - Italy
Email: cicirelli@icar.cnr.it, spezzano@icar.cnr.it

KEYWORDS

Sensor data fusion; Internet of Things; Multi-agent systems; Statecharts; MAB museum.

ABSTRACT

Sensor data fusion refers to technological solutions aiming at collecting, classifying and complementing data coming from multiple sensors. It has the potential of enabling context awareness which, on the other hand, represents a huge potential to be exploited in the field of IoT applications. Sensor fusion and IoT have to deal with multi-faced issues like heterogeneity, sensor/actuator management, data accuracy and reliability. This paper proposes a multi-tier approach dealing with sensor fusion and IoT aspects in a modular way. The approach relies on the use of the agent metaphor, statecharts and on the Rainbow multi-agent platform. Agents can be dynamically added and removed from an application thus promoting system openness and scalability. Heterogeneity and distribution issues are transparently managed by Rainbow which hides the physical layer on top of which the applications are built. As a significant case study, the approach was exploited for the implementation of a working prototype devoted to improve security of some artworks (statues) of the MAB museum located in the city of Cosenza, Italy.

INTRODUCTION

Nowadays, sensors are exploited in a huge variety of applications ranging from healthcare, transportation and logistic, surveillance, environmental monitoring, smart city and so forth. The sensing capabilities of the infrastructures and devices surrounding our daily lives are constantly improving and becoming more affordable. The widespread diffusion of sensors was also favored by the ever-increasing attention which is given to the field of the so called "Internet of Things" (IoT) (Miorandi et al., 2012). The basic idea of the IoT is a pervasive presence around us of a variety of things or objects such as RFID tags, sensors, actuators, mobile phones, etc. which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbors to reach common goals (Atzori et al., 2010).

Important issues which are related to the development of significant sensor-based applications are data collection, classification and complementation. Sensor data fusion, or simply *sensor fusion* (SF) (Karimi, 2013), refers to technological solutions having the aim of addressing the above issues. The

goal is to increase both the accuracy and reliability of sensed data as well as to enable context awareness (Bicocchi et al., 2014; Karimi, 2013; Schilit et al., 1994). Context awareness refers to the capability of disclosing information about the context (or situation) in which the data get acquired/generated. This allows to validate an event or an assumption made on sensed data, or to compensate the lack of complete information about the sensed environment thus permitting to take decisions and/or properly react to complex environmental stimuli. As an example, a person may not see flames under the hood of a car, but the smell of burning rubber and the heat coming from the dash would suggest that it is prudent to leave the car because the engine is on fire.

This paper proposes an approach for the development of distributed SF applications based on the IoT paradigm. The approach promotes separation of concerns through the use of a *multi-tier architecture* which allows to deal with SF and IoT issues in a modular and orthogonal way. The agent metaphor (Woolridge and Wooldridge, 2001) is used to structure the application logic, whereas agents' behavior is modeled by using statecharts (Booch et al., 2000; Cicirelli et al., 2011; Harel, 1987; Kielar et al., 2014). Statecharts are a state-based formalism which allows to specify complex and time-dependent behaviors by using a graphical notation. Complexity of a model is dealt with the use of hierarchical constructs.

Proposed approach permits the exploitation of *concept hierarchies* which allow the *classification of high-level situation* as well as the definition of *reaction-driven policy* which work on a multiplicity of low-level events. Concept hierarchies are useful to represent a system as a set of abstractions normally used by humans when reasoning about complex systems. The concepts which are introduced in different tiers, or within the same tier, can be related together (i.e., orchestrated) in order to specify complex reaction patterns to the stimuli coming from the external environment. The overall goal is to move from the IoT towards the Cognitive IoT (Tsai et al., 2014; Wu et al., 2014) where objects interact and operate by acquiring knowledge from the surrounding environment and by following a context-aware perception-action operational cycle.

Each tier, in the proposed architecture, groups agents on the basis of some predefined agent roles. For instance, the role *awareness* was defined, and the corresponding tier will contain agents devoted to managing the knowledge acquired on the whole system and to properly act on the system itself. A *low-level* tier, instead, contains objects (not necessarily agents)

used to deal with the hardware/driver system entities.

The agent metaphor was chosen for its capability to naturally meet some needs of distributed SF applications like the autonomy in reacting to stimuli coming from the external environment, and its suitability to operate in an open and dynamic environment. Heterogeneity and distribution issues are transparently managed by the Rainbow (Giordano, Spezzano and Vinci, 2014) agent platform which has the capability of abstracting the physical layer on top of which the applications are built.

The proposed approach is novel because it is *comprehensive*, and explicitly considers aspects tied to sensors and actuators. Comprehensive means that it takes into account all the issues relevant to the development of SF applications starting from the awareness of the context down to the issues related to the management of physical layer. Flexibility and modularity is provided by the tiered-based architecture and by the use of the agent metaphor paired with statecharts.

As a significant case study, the approach was exploited for the implementation of a working prototype devoted to improve security of some artworks (statues) of the open-air MAB museum located in the city of Cosenza, Italy.

The paper is structured as follows. First, the proposed approach is described along with a characterization of the proposed multi-tier architecture. Then, an overview of the Rainbow agent-based platform is provided. After that, the case study is shown. Conclusions and indications about future research avenues are provided in the last section.

A MULTI-TIERS APPROACH FOR SF

The proposed multi-tiers approach allows the development of SF applications by promoting both separation of concerns and modularity. In each tier it is possible to focus on specific concerns having a different level of abstraction. The *low-level* tier, for instance, copes with software drivers managing physical devices. The *high-level* one deals with context/situation awareness. The terms *low* and *high* refer to the abstraction level exploited for developing an application or a working system. The use of tiers fosters both a *top-down* and *bottom-up* software development schema. A top-down approach can be used when an application is developed from scratch. A bottom-up approach, instead, is suggested to be used when an application is built on top of some preexisting components. The introduced tiers are reported in Figure 1. All the tiers except the lower one, are populated by means of agents. The *low-level* tier instead contains objects which are referred as *Virtual Objects* (VOs). Such name mirrors the fact that VOs are used to virtualize, i.e., to abstract, the physical devices used for realizing an application. A description of the introduced tiers follows.

- *Low-level*: it hosts the VOs that directly interact with physical devices. In the case of simple devices, such virtual objects directly manage information like distances, level of noise or temperature (both for sensing and actuation purposes). In other more complex cases, instead, virtual objects interact with the drivers of the physical devices and can abstract complex operations (e.g., those related to the management of a webcam).

The goal of this tier is to decouple, from a sensing/actuation point of view, the functionalities from the equipment offering them. A change in the physical devices affects virtual objects but it should not affect the entities belonging to the other (upper) tiers. The challenge, here, is to decouple *functionalities* from *implementation*. A one-to-many mapping exists between VOs and the hardware equipments.

- *Sensing/Actuator*: the agents defined in this tier are boundary entities whose goal is turning the functionalities offered by virtual objects into location-independent services offered by agents. As a general guideline, it is suggested to develop a dedicated agent for all the functionalities offered by a single device abstracted through a VO. These dedicated agents must be co-located on the computational node hosting the virtual objects they have to interact. From a functional point of view, the Sensing/Actuator agents do not introduce new functionalities nor modify the existing ones. They can instead mediate the use of the wrapped functionalities by enforcing, for instance, *negotiation procedures* and/or *access policies* useful for guaranteeing an exclusive use, or a time-based use, of such functionalities.
- *CSensing/CActuator*: here the goal is to compose, or modify, functionalities offered by the agents belonging to the Sensing/Actuator tier. Composition of functionalities aims at obtaining aggregate operations starting from simpler ones, and at orchestrating independent functionalities in order to pursue a common task. Modify a functionality means to extend/improve it, e.g., for increasing the reliability of sensed data, or for allowing the transparent management of a group of redundant actuator devices viewed as a single device. As an example, if it is required to average the data coming from multiple temperature sensors, it is possible to consider an *averaging agent* which interacts with the agents related to the temperature sensors in order to get data and average it. In another case, the functionality offered by a temperature agent which provides only data in real-time, can be extended by an agent providing information about the maximum and minimum value of read data. Modify a functionality can mean also to hide it in order to meet specific application constraints. A many-to-many relationship can exist between agents in this tier and the lower one.
- *Concept*: this tier introduces a *significant increase* in the exploitable abstraction level. In fact, in the tiers previously described we have been dealing only with issues directly related to sensing and controlling the environment, e.g., reading a temperature value or closing a light. In this tier, instead, an actuation to be performed, and/or some environmental stimuli, are mapped onto more abstract concepts. Each concept is then modeled/implemented through an agent. The goal, here, is to allow reasoning by using abstract concepts directly related to the specific domain of the application being developed. For instance, in this tier, it is possible to introduce the concept of *climatic health* which depends on the value of the temperature

and humidity actually existing in a given room. In the same way, the concept of *office security* can be introduced in order to control a door office and makes it alarmed after a certain hour. A many-to-many relationship can exist between agents in this tier and the lower one.

- *CConcept*: this tier is equivalent to the CSensing/Cactuator one. In this case, though, composition and modification refer to concepts.
- *Awareness*: this is the tier permitting to operate, and to reason, at the highest level of abstraction. Agents introduced in this tier have a holistic vision of all the concepts which are related to the developed application and which govern the implemented system. Introduced concepts (or a subset of them) are orchestrated in order to pursue application goals. Information and data can be complemented in order to augment the knowledge about the context in which the application runs. For instance, if an application for managing *safety at home* is considered, information indicating that a person is found lying in the bathroom can be complemented in order to infer that the person could be afflicted by a malaise.

Figure 2 reports an example in which a certain number of agents and VOs populate the various tiers and interact among them. In the figure it is highlighted that communication among the introduced entities can be both unidirectional and bidirectional. Even if in the reported example all the tiers contain some entities, on the bases of application needs, a tier could also be empty. It is not allowed to consider intra-tier communication among the entities located in the tiers from T#3 to T#0 whereas it can occur, instead, in tiers T#4 and T#5. All of this favors to deal with the *true* application logic (e.g., the goal-oriented logic) only in the tiers with a higher level of abstraction.

Once the tiers are populated, and having established the communication patterns among the introduced entities, the next step is to define the behavior of each agent. In the approach here proposed, modeling the behavior of agents relies on the use of statecharts (Booch et al., 2000; Cicirelli et al., 2011; Harel, 1987; Kielar et al., 2014). As a consequence, the whole system can be seen as a network of interacting statecharts. Statecharts are well suited to model entities having a complex and time-dependent behavior. They enable both a hierarchical and modular modeling approach. All of this permits to face with the well-known state-explosion phenomenon which arises with large and complex models. The basic mechanism upon which statecharts rely, consists in the possibility of nesting a subautomaton within a (macro) state thus encouraging step-wise refinement of a complex behavior.

THE RAINBOW PLATFORM

Rainbow (Giordano, Spezzano and Vinci, 2014; Giordano, Spezzano, Vinci, Garofalo and Piro, 2014) is a distributed agent-based IoT platform composed of single-board computer nodes, such as the Raspberry PI 2, which is well suited to connect massive-scale networks of sensors and actuators to the Cloud. It was chosen as the reference platform exploitable for supporting the approach so far described.

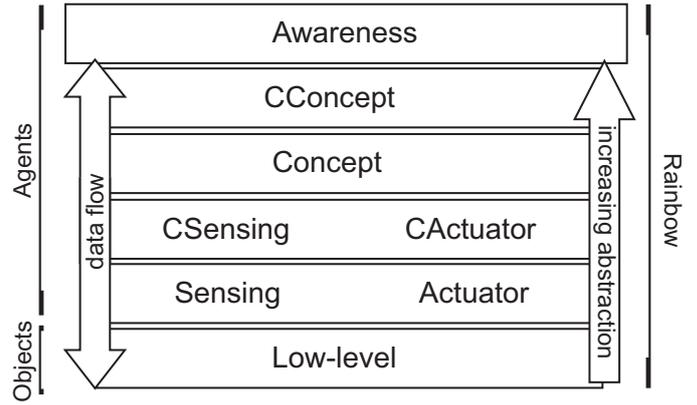


Fig. 1. The proposed multi-tier architecture

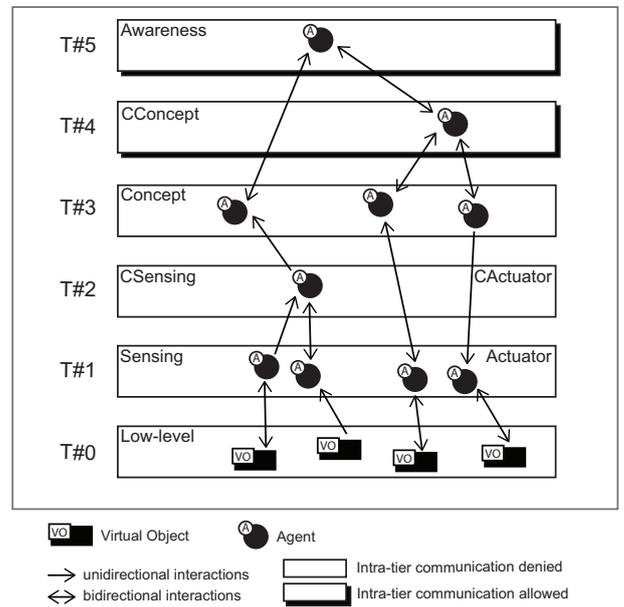


Fig. 2. The multi-tier architecture: an example of communication details

The physical part of the Rainbow architecture is constituted of sensors and actuators, together with their relative computational capabilities, which are directly immersed in a physical environment. Physical entities are usually spread across a large (even geographic) area. All of this implies that the controlling part of an application must be intrinsically distributed. Sensors and actuators are partitioned into groups, each of which is managed by a single computing node. A goal of Rainbow is to bring the computation (e.g., the controlling part) as close as possible to the physical part. Each computing node hosts multi-agent applications designed to monitor multiple conditions, or to operate activities within a specific environment. For this purpose, each node contains an *agent server* that permits agents to be executed properly. Agents can be intelligently assisted by cloud services, that support complex analytics, modeling, optimization and visualization tools, to make better operational decisions.

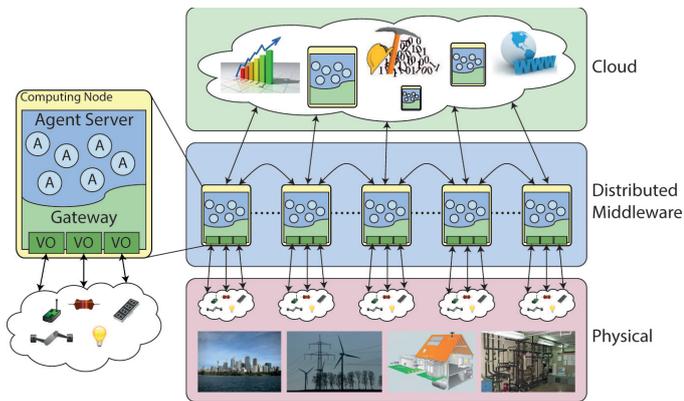


Fig. 3. The Rainbow architecture

The architecture of Rainbow (see Figure 3) is composed of the Cloud layer, the Intermediate layer and the Physical layer.

In the Intermediate layer, sensors and actuators of the Physical layer are represented as virtual objects (VOs). VOs offer to the agents a transparent and uniform access to the physical part through the use of a well-established interface. A VO allows agents to connect directly to devices without care about proprietary drivers and without addressing some kind of fine-grained technological issues. Each VO comprises *functionalities* directly provided by the physical part. Essentially, a VO exposes an abstract representation (i.e., a machine-readable description) of the features and capabilities of the abstracted physical objects spread in the environment. Functionalities exposed by different types of VOs can be combined in a more sophisticated way on the basis of event-driven rules which affect high-level applications and end-users. More in particular, all the devices are properly wrapped in VOs which, in turn, are enclosed in distributed *gateway containers*. The computational nodes that host the gateways and the agent server represent the middle layer of the Rainbow architecture.

Gateways and agent servers are co-located in the same computing nodes in order to guarantee that agents directly exploit the physical part through VO abstraction. Instead of transferring data to a central processing unit, we actually transfer the process (i.e., the fine-grain agent's execution) toward the data sources. As a consequence, less data needs to be transferred over a long distance (i.e., toward remote hosts) and local access and computation will be fostered in order to achieve good performance and scalability.

The upper layer of Rainbow architecture concerns the cloud part. This layer addresses all the activities that cannot be properly executed in the middle layer like, for instance, algorithms needing a complete knowledge of the whole system, tasks that require high computational resources, or when a historical data storage is mandatory.

SENTIENT STATUES IN THE MAB MUSEUM

The MAB is a particular open-air museum located on the main artery of the new part of the city of Cosenza (Calabria, southern Italy). The MAB was born thanks to the donation of the wealthy collector Carlo Bilotti, native to Cosenza but immigrant to America, who decided to donate his art

collection to his city of birth after his death in 2006. The MAB houses some prestigious sculptures by artists like Salvador Dalí, Giorgio De Chirico and by some other artists of Calabria. Since its establishment, the museum has been the subject of some vandalisms and of some accidents that damaged some artworks.

The goal of the case study which is here described, is to use sensor fusion concepts in order to *furnish* to the statues of the MAB some *virtual senses* able to *make them aware* of what happens around them. The aim is therefore to make the statues *sentients*, i.e. able to discover and recognize a dangerous situation and to react to it with the aim of preventing damages and averting a possible hazard. The basic idea is to equip a sculpture with some suitable sensors and actuators enabling to sensing the environment and operate deterrence actions, e.g. asking for help. Obviously, in order to avoid causing damage to the artworks, both sensors and actuators should not be *worn* by the sculpture but deployed in their vicinity in some safe places.

A prototyped system was developed by using the approach described in this paper. The system was developed from scratch and, for this reason, a top-down development schema was exploited. More in particular, in a first phase, the layers of the multi-tiers architecture were populated starting from the *awareness* layer. Then, the behavior of each agent was modeled through statecharts and the code for the modeled agents and virtual objects was developed. Finally, the system was implemented. System implementation consisted in preparing the hardware equipment (i.e., the chosen sensors and actuators), connecting it to a Rainbow server and deploying the developed code of both agents and virtual objects over that server. The simplest devices were connected to the Rainbow server through some Arduino devices.

In the following, a description of the developed prototype is reported. A description of the basic features of statecharts is provided too.

Populating the layers of the multi-tiers architecture

This is the most important phase because it defines the roles and the functionalities of all the entities which constitute the system. In Figure 4 are reported the agents and the VOs defined for the application.

In the tier of *awareness* we defined the agent which models the *mood* of a sculpture. On the base of the stimuli received from the environment, a statue can find itself in a status like quiet, worried or terrified. A detailed description of these status is provided in the next subsection. A statue perceives its surrounding environment through its *senses*. Three different senses were considered, namely the *touch*, the *sight* and the *hearing*. For each of the considered senses, a specific agent was defined in the *concept* layer.

On the base of the current status, some *safeguard* actions can be executed. For this purpose, a suitable agent is defined in the *concept* tier. This safeguard agent models a composite concept which is tied to the concepts of both *speaking* and making *deterrent actions* (see the Speech and Deter agents in the *concept* layer of Figure 4).

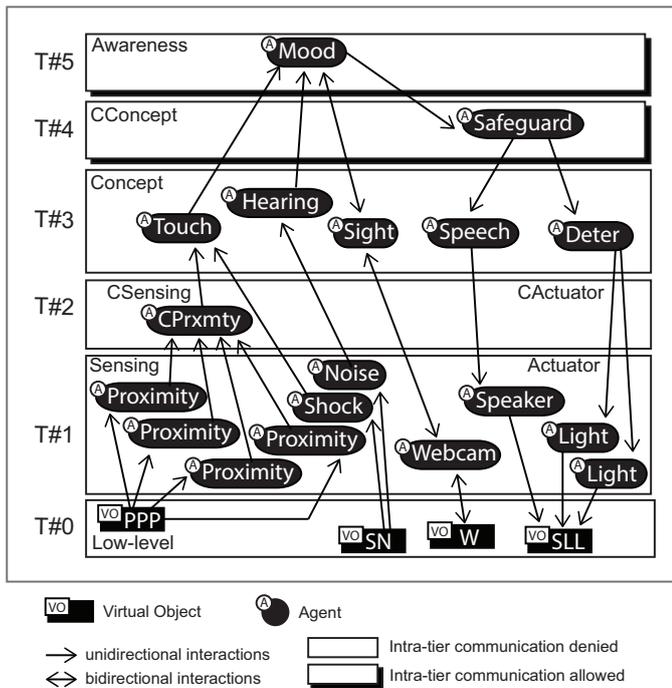


Fig. 4. Realizing sentient statuses: agents' hierarchy and virtual objects

From the *sensing* tier it emerges that (i) the sense of touch is implemented by using both some proximity sensors and a shock sensor, (ii) the hearing is implemented by a noise sensor and (iii) the sight is realized by using a webcam. Proximity sensors are grouped in a single composite sensor (see the CPrxmt agent in the csensing layer of Figure 4). The latter composite entity was considered to highlight that, in this application, all the proximity sensors are indistinguishable and equivalents.

From the actuator point of view, the safeguard activities were implemented by considering some speech abilities realized through a speaker, and by some deterrent actions which are based on the flashing of warning lights.

Four VOs were considered: one for the management of proximity sensors, one for the shock and noise sensor, and another two respectively for the webcam and the warning lights.

Basic features of statecharts

A *state* of a statechart can recursively be decomposed into a set of *substates*, in which case such a state is said to be a *macro* state. A state that is not decomposed is said to be a *leaf* state. The root state of the decomposition tree is the only one having no parent and it is referred to as the *top* state. At a given point in time, a statechart finds itself simultaneously in a set of states that constitutes a path leading from one of the leaf states to the top state. Such a set of states is called a *configuration* (Harel, 1987). A configuration is uniquely characterized by the only leaf state which it contains. Each macro state specifies which of its substates must be considered its initial state. This substate is indicated by means of a curve originating from a small solid circle and ending on its border. This curve, although technically

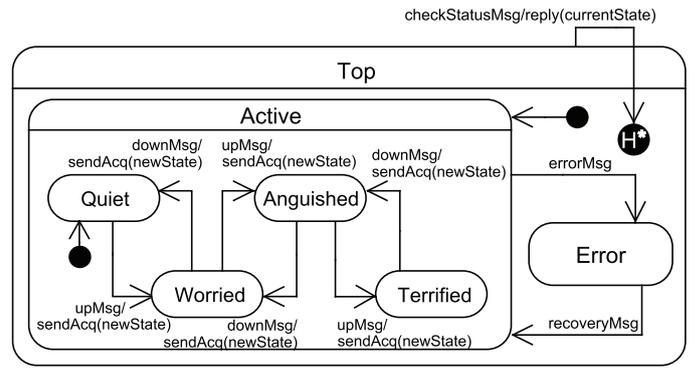


Fig. 5. Statechart of a Mood agent for a sentient statue

is not a transition, is referred to as the *default transition*. State transitions are represented by edges with arrows. Each transition is labelled by $ev[guard]/action$ where ev is the trigger (event or message causing the transition), $guard$ a logical condition which enables the transition when it evaluates to true, and action the action “à la Mealy” associated with the transition. When omitted, the guard is implicitly assumed to be true. Both source and destination of a transition can be states at any level of the hierarchy. Firing a transition leads the statechart to switch from one configuration to another. When a configuration is left, each of its macro states keeps memory of its direct *substate* that is also part of the configuration. This *substate* is referred to as the *history* of the macro state.

A transition always originates from the border of a state, but it can reach its destination state either on its border or ending on a particular element called *history connector* or H-connector. Such a connector is depicted as a small circle containing an H (*shallow history*) or an H* (*deep history*), and it is always inside the boundary of a compound state. Both shallow and deep history connectors allow to re-enter in a macro state by exploiting history information.

Modeling agents through statecharts

The behavior of each agent identified in Figure 4 was modeled through statecharts. For simplicity, here, only the statechart of the *mood* agent is described. The statechart is reported in Figure 5.

The mood agent can find itself in the *Active* or *Error* state. The above two status are contained in the *Top* macro-state having the responsibility or responding to *check* messages asking for the current status of the statue. Each time one of such a message is received, the statue replies with its actual state and then returns in the leaf-state owned just before receiving the check message. All of this is mirrored in the model by the use of the deep-history connector H* which is attached to the state-transition edge departing from the *Top* state.

The *Active* state is also a macro-state. It contains the leaf-states related to the actual “morale” of a statue, namely *Quiet*, *Worried*, *Anguished* and *Terrified*. A mood agent moves from a status to another one by receiving *upMSG* and *downMSG* messages. These messages are generated by the senses of the statue. For instance, in the case the *touch* identifies the presence of someone in the nearness of the statue, an *upMSG*

```

[12/11/2015 12:15:13] Mood#24: received UpMsg, current state QUIET,
next state WORRIED. Send WORRIED to: Safeguard#12, Sigh#5
[12/11/2015 12:20:21] Mood#24: received DownMsg, current state WORRIED,
next state QUIET. Send QUIET to: Safeguard#12, Sigh#5
[12/11/2015 12:21:42] Mood#24: received UpMsg, current state QUIET,
next state WORRIED. Send WORRIED to: Safeguard#12, Sigh#5
[12/11/2015 12:22:12] Mood#24: received UpMsg, current state WORRIED,
next state ANGUISHED. Send ANGUISHED to: Safeguard#12, Sigh#5
[12/11/2015 12:23:19] Mood#24: received UpMsg, current state ANGUISHED,
next state TERRIFIED. Send TERRIFIED to: Safeguard#12, Sigh#5
[12/11/2015 12:26:29] Mood#24: received CheckMsg, current state
TERRIFIED. Send TERRIFIED to: OtherAgent#7
[12/11/2015 12:26:45] Mood#24: received DownMsg, current state TERRIFIED,
next state ANGUISHED. Send ANGUISHED to: Safeguard#12, Sigh#5
[12/11/2015 12:27:45] Mood#24: received UpMsg, current state ANGUISHED,
next state TERRIFIED. Send TERRIFIED to: Safeguard#12, Sigh#5
[12/11/2015 12:28:02] Mood#24: received DownMsg, current state TERRIFIED,
next state ANGUISHED. Send ANGUISHED to: Safeguard#12, Sigh#5
[12/11/2015 12:28:45] Mood#24: received DownMsg, current state ANGUISHED,
next state WORRIED. Send WORRIED to: Safeguard#12, Sigh#5
[12/11/2015 12:29:10] Mood#24: received CheckMsg, current state
WORRIED. Send WORRIED to: OtherAgent#7
[12/11/2015 12:29:25] Mood#24: received DownMsg, current state WORRIED,
next state QUIET. Send QUIET to: Safeguard#12, Sigh#5

```

Listing 1: An example of generated log

TABLE I. DESCRIPTION OF THE STATUE STATUS

Status	Description
Quiet	The data coming from the sensors gets elaborated. The webcam is turned on with a fixed shot on the statue. The lights and the speakers are switched off.
Worried	The data coming from the sensors get elaborated. The webcam is turned on with a movable frame around the statue. The lights and the speakers are switched off.
Anguished	The data coming from the sensors gets elaborated. The webcam is turned on with a movable frame around the statue. The lights are flashing and the speakers are switched off.
Terrified	The data coming from the sensors gets elaborated. The webcam is turned on with a movable frame around the statue. The lights are flashing and the speakers invited to move away from the statue
Error	The data coming from the sensors is discarded. The webcam, the lights and the speakers are switched off.

message is issued. On the contrary, when a previously sensed entity is no longer sensed, a *downMSG* message is sent to the mood agent. In such a way, a statue stays in the *Quiet* state when no senses reveal the presence of anyone in the nearness of the statue. On the contrary, when the senses get stimulated all together, the statue moves to the *Terrified* status. When a statue changes its status, a *newState* message is sent to the acquaintances of the statue. These messages carry out information about the new reached state. All the leaf-states of a statue, along with their description, are reported in Table I.

Description of the realized prototype

Figure 6 portrays the realized prototype. It consists in a thin wooden support upon which we arranged the lights, the speakers and a plastic case hosting: the Rainbow server running over a Raspberry device, the Arduino used to manage both the sensors and the actuators, the noise sensor and the shock sensor. The proximity sensors were instead fixed under the wooden support. A placeholder of the statue was placed over a transparent container hosting the lights. The webcam was instead attached to a metal support placed in the nearness of the equipment so far described.

The arrangement of the whole equipment takes into account how the system can be really placed in the museum. The statues, in fact, are placed over a large plexiglass base. The prototyped system was conceived to be embedded in such base

with the lights, when flashing, visible from the outside of the base. The webcam was instead expected to be installed on the nearest building with respect to a statue.

In Listing 1 is reported an excerpt of the log file produced by a Mood agent during system execution. From the log it is possible to see how the agent reacts to the event coming from its senses and how it change its status accordingly.

CONCLUSIONS

This paper has presented a multi-tier approach for developing SF applications in the domain of the IoT. The approach is based on the use of the agent metaphor, statecharts and on the Rainbow multi-agent platform. Agents naturally allow the development of distributed applications in a open and dynamic environment. Statecharts are well suited to model entities having a complex and time-dependent behavior. Model complexity is dealt with modular and hierarchical constructs. The Rainbow platform has the capability of hiding the physical layer on top of which the applications are built.

Different levels of abstractions can be exploited whilst designing an application. All of this fosters modularity and separation of concerns. Both a top-down and a bottom-up development schema can be adopted. The paper furnishes some guidelines for using the provided abstract levels. The goal is to replace the management of the low-level environmental stimuli and actuations to be performed with some high-level concepts which are close to the considered application domain.

As a significant case study, the approach was exploited for implementing a prototype aiming at improving security of some artworks (statues) which are in the MAB museum located in the city of Cosenza, Italy.

Prosecution of the work is devoted to:

- making the realized prototype really working in the museum
- allowing the statues to cooperate with each other in order to discover and prevent dangerous situations
- extending the functionalities of the provided prototype by offering services for increasing usability of the museum (e.g., for making virtual tour in the museum or accessing to a virtual bulletin board) and by allowing the use of the senses of the statue on behalf of tourist needs (e.g., by using the sight for taking a selfie)
- making available a tool for the automatic generation of agent-code starting from the modeled statecharts
- enhancing the approach so as to exploit not only modeling capabilities but also making possible the analysis of the realized models (e.g., through distributed simulation (Cicirelli and Nigro, 2013)) before their final implementation.

ACKNOWLEDGMENT

This work has been partially supported by RES-NOVAE - "Buildings, roads, networks, new virtuous targets for the Environment and Energy" project, funded by the Italian Government (PON 04a2_E).

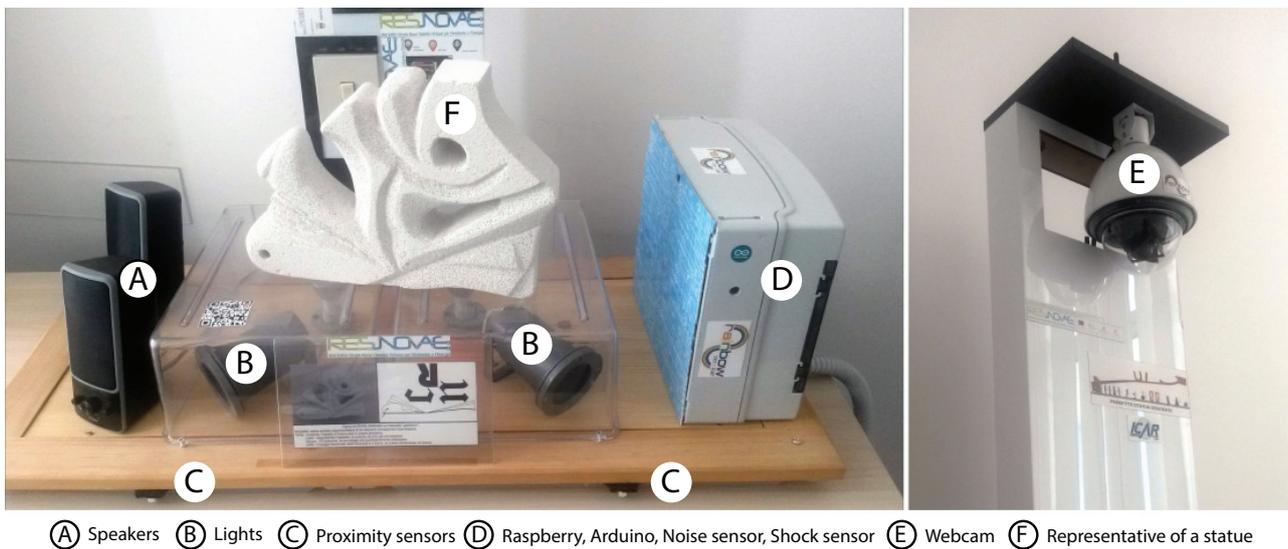


Fig. 6. Picture of the realized prototype

The authors wish to thank Christian Nigro, Alessandro Mercuri, Elisa Coscarella and Emilio Greco for their valuable collaboration in the development of the work here proposed.

REFERENCES

- Atzori, L., Iera, A. and Morabito, G. (2010), The internet of things: A survey, *Computer networks* 54(15), 2787–2805.
- Bicocchi, N., Fontana, D. and Zambonelli, F. (2014), A self-aware, reconfigurable architecture for context awareness, *IEEE Symposium on Computers and Communications, ISCC 2014*, Funchal, Madeira, Portugal, June 23–26, 2014, pp. 1–7. <http://dx.doi.org/10.1109/ISCC.2014.6912485>
- Booch, G., Rumbaugh, J. and Jacobson, I. (2000), *The Unified Modeling Language User Guide*, Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.
- Cicirelli, F., Furfaro, A. and Nigro, L. (2011), Modelling and simulation of complex manufacturing systems using statechart-based actors, *Simulation Modelling Practice and Theory* 19(2), 685–703. <http://dx.doi.org/10.1016/j.simpat.2010.10.010>
- Cicirelli, F. and Nigro, L. (2013), *Communications in Computer and Information Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter An Agent Framework for High Performance Simulations over Multi-core Clusters, pp. 49–60.
- Giordano, A., Spezzano, G. and Vinci, A. (2014), Rainbow: an intelligent platform for large-scale networked cyber-physical systems, *Proceedings of the 5th International Workshop on Networks of Cooperating Objects for Smart Cities (UBICITEC 2014)* co-located with CPSWeek 2014, Berlin, Germany, Apr 14, 2014., pp. 70–85.
- Giordano, A., Spezzano, G., Vinci, A., Garofalo, G. and Piro, P. (2014), A cyber-physical system for distributed real-time control of urban drainage networks in smart cities, *Internet and Distributed Computing Systems*, Springer, pp. 87–98.
- Harel, D. (1987), Statecharts: A visual formalism for complex systems, *Sci. Comput. Program.* 8(3), 231–274.
- Karimi, K. (2013), The role of sensor fusion and remote emotive computing (rec) in the internet of things, Freescale Semiconductor. http://cache.freescale.com/files/32bit/doc/white_paper/senfeiotlfpw.pdf
- Kielar, P. M., Handel, O., Biedermann, D. H. and Borrmann, A. (2014), Concurrent hierarchical finite state machines for modeling pedestrian behavioral tendencies, *Transportation Research Procedia* 2, 576 – 584. <http://www.sciencedirect.com/science/article/pii/S2352146514001343>
- Miorandi, D., Sicari, S., De Pellegrini, F. and Chlamtac, I. (2012), Internet of Things, *Ad Hoc Netw.* 10(7), 1497–1516. <http://dx.doi.org/10.1016/j.adhoc.2012.02.016>
- Schilit, B. N., Adams, N. and Want, R. (1994), Context-aware computing applications, *Workshop on Mobile Computing System and Applications*, IEEE Computer Society, pp. 85–90.
- Tsai, C.-W., Lai, C.-F. and Vasilakos, A. V. (2014), Future internet of things: Open issues and challenges, *Wirel. Netw.* 20(8), 2201–2217.
- Woolridge, M. and Wooldridge, M. J. (2001), *Introduction to Multi-agent Systems*, John Wiley & Sons, Inc., New York, NY, USA.
- Wu, Q., Ding, G., Xu, Y., Feng, S., Du, Z., Wang, J. and Long, K. (2014), Cognitive internet of things: A new paradigm beyond connection., *IEEE Internet of Things Journal* 1(2), 129–143.

A SIMULATION BASED STUDY OF THE EFFECT OF TRUCK ARRIVAL PATTERNS ON TRUCK TURN TIME IN CONTAINER TERMINALS

Ahmed E. Azab and Amr B. Eltawil

Department of Industrial Engineering and Systems Management
Egypt-Japan University of Science and Technology (E-JUST)
POBox 179, New Borg Elrab city, 21934 Alexandria, Egypt
E-mail: ahmed.azab@ejust.edu.eg, eltawil@ejust.edu.eg

KEYWORDS

Container Terminal, Truck Arrival Pattern, Truck Turn Time, Discrete Event Simulation, Truck Appointment System.

ABSTRACT

In container terminal operations, the delay of containers delivery is a common problem that confronts both the terminal operator and the customers represented by the trucking companies. One source of this delay is due to the long waiting time of the transporting trucks at container terminals (CTs). In this paper, the problem of long turn time for external trucks is studied. An extensive review of the previous work available in the literature that focused on landside problems in CTs is presented. After identifying some gaps, we conclude that the arrival pattern of external trucks and its impact on the truck turn time needs to be more understandable. A discrete event simulation model is developed to study the effect of various truck arrival patterns on the truck turn time in CTs. The simulation results showed how the arrival patterns influence the turn time of external trucks. Moreover, we suggest how CT operators can reduce the turn times without reducing the terminal gates' productivity and recommend how to consider the arrival pattern in designing an appointment system for external trucks in CTs.

INTRODUCTION

Container terminals (CTs) received considerable interests from researchers in recent years. This is due to the importance of container terminals as essential nodes in global supply chains. The global trade growth put

more pressures on CTs to manage its activities and schedule its resources properly. These growing activities increased the complexity of CT related planning and operational control problems. By solving CT's problems efficiently, the terminal puts its position on the map of global competitiveness among other container terminals. The increasing number of containers causes higher demands on the seaport container terminals, container logistics, and management, as well as on technical equipment (Steenken et al. 2004). As a result, CTs are forced to enlarge handling capacities and strive to achieve gains in productivity (Stahlbock and Voß 2008).

Container terminals can be divided into five areas, namely the berth, quay, transport, storage yard, and gate, as illustrated in Fig. 1 (Carlo et al. 2013). These five areas can be described in three main areas as follows: the berth and the quay are called the "Seaside", the yard storage is "yard side" while the gate is called the "Landside". Each side has some operations which interact with others (Fig. 2). Sometimes, these interactions are studied as integrated problems. Solving the integrated problems has many benefits. Karam et al. (2014) showed that the integration achieves the required performance of each individual problem and gives better solutions.

The yard is considered the center area of any CT. As a result, yard operations interact with both seaside operations and landside operations. Fig. 2 describes the main areas of the terminal and the operations interactions. The main seaside operations are: berth allocation for vessels, quay crane allocation, handling containers from/to the vessel using quay cranes, and

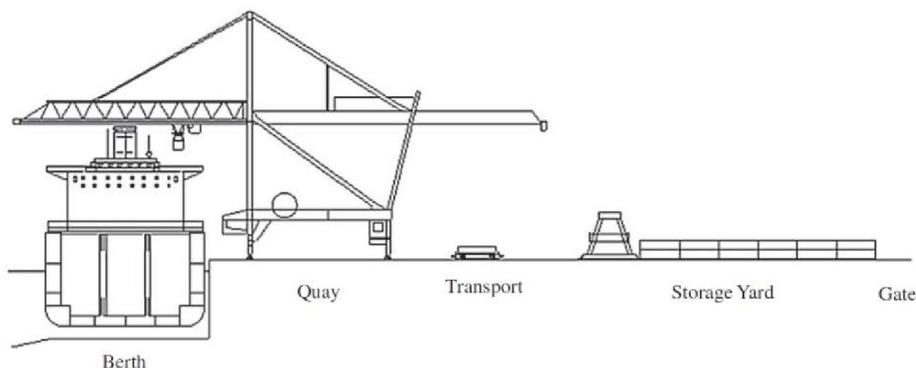


Fig. 1 Container Terminal Main Areas (Carlo et al. 2013).

loading/unloading containers to/from the internal transport means. Internal trucks are delivering containers between seaside and yard. The yard area main operations are: loading/unloading of containers to/from internal and external trucks, and stacking containers in the appropriate locations in the yard, as well as premarshalling of containers. To transport containers between the terminal and the hinterland, external trucks can access the terminal via the terminal gates. The main landside operations are described in the next section.

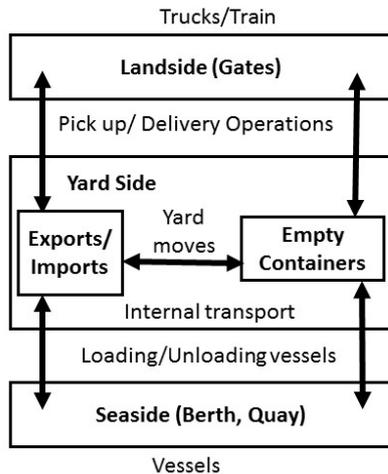


Fig. 2 Operation areas of a seaport container terminal and transport flows

Seaside operations and yard operations received abundant interest from researchers while landside operations still need more efforts to tackle problems in this important area of the CTs. In this paper, focus will be more on the land side operations. Various problems will be addressed, solution strategies and approaches for the landside operations from the literature will be shown with more attention to the use of simulation as a powerful solution methodology for this class of problems.

The remainder of this paper is organized as follows. The next section will discuss operations and problems in the landside. After that an extensive literature review for landside operations will be introduced. The problem description follows, with an explanation of the used simulation model and the experimental results. Conclusions and future work will be addressed last.

LANDSIDE OPERATIONS AND PROBLEMS

The landside is considered the connecting point between the terminal and the hinterland. When a terminal receives an announcement for vessel(s) arrival, the terminal announces to the customers. Once the customers know the expected arrival schedule of the vessel(s), they make orders to trucking companies or their own trucks to deliver or pickup their containers. In some seaports, trucking companies have to make an

appointment based on the expected arrival of the vessel and the terminal schedules. Some trucks may go to the terminal without an appointment. At the terminal gates, the trucker's identity, the truck's documents and the container's documents are checked. Export/Import containers are scanned before entering/leaving the terminal. When these operations are finished, the truck is directed to the yard. The gate operator gives the trucker the precise location and the route that he/she should take to reach the container location at the storage block.

At the yard, a yard crane will load the import container to the truck or unload the export container from the truck. The handling operation of the yard crane is also scheduled. CTs often separate the storage locations of export blocks and import blocks. All previous operations need to be well managed and planned in the long term and scheduled in the short term to guarantee high productivity and service quality level for the terminal.

Problems in the landside can be studied from two perspectives. The first is the customer perspective, and the second is the terminal perspective. In CTs, the operations in the landside are interacting with the yard operation (see Fig.2), and the main objective of landside operations is to serve the external trucks at the gates and the yard. The terminal main goals are: increasing the utilization of the terminal equipment and facilities and achieving high customers' satisfaction. Customer's satisfaction occurs when they are served at minimal time and cost. The more time the trucks wait at the gates and the yard, the more queuing problems will occur. As a result, the long queues of trucks at the terminals lead to delays and cause emissions, congestion, and high cost for both the terminal and the customer (Gharehgozli et al. 2014). More focus on the delay problem for external trucks will be addressed in this paper.

Due to the complex interactions of processes in CTs, simulation has been used to solve many CT problems. Moreover, Control techniques which relate to the dynamic behavior of the equipment are even more difficult to analyze and benchmark, therefore necessitating the presence of a tool that can replicate the behavior of a real terminal (Angeloudis et al. 2011). It is difficult to predict the actual events such as arrival times of trucks and the actual number of arrivals. Simulation is very effective in doing such predictions. On the other hand, it is very important to study the "what if" scenarios to take the right decisions in the short-term and long-term.

LITERATURE REVIEW

In this section, the published papers after 2008 that focused on landside operations are reviewed with more attention for the trucking problems at gates and yard.

Simulation based work to solve landside problems is also covered.

Truck operations in the landside are evaluated using some performance metrics like “Truck turn times”. Both terminal and truckers wish to minimize the truck turn time. The turn time is defined as the time from the truck arrival at the terminal gates to the time of departure. Huynh (2009) provided a mathematical model to examine the effect of limiting truck arrivals on truck turn time and crane utilization. To obtain the average truck turn time, the authors used a discrete event simulation (DES) model and heuristics to solve their model. As an extension for their work, Huynh and Walton (2011) produced DES model to simulate various appointment rules. They examined the individual appointments vs the block appointment and studied its effect on truck turn time and crane utilization. In a previous work, Huynh and Hutson (2008) examined the sources of delay for dray trucks at container terminals. They used decision trees as a powerful data mining tool.

To reduce the transportation cost, Namboothiri and Erera (2008) proposed an integer programming model and solved it using heuristics. The solution provided the best set of appointment reservations and routes for a fleet of trucks. Guan and Liu (2009) formulated a non-linear optimization problem and applied a multi-server queuing model to analyze marine terminal gate congestion and quantify truck waiting cost. They found that the truck appointment system seems to be the most viable way to reduce gate congestion and increase system efficiency. An optimization model for truck appointments is formulated by Zhang et al. (2013) to reduce heavy truck congestion in the terminal. The queuing process described by a Baskett Chandy Muntz Palacios (BCMP) queuing network. To solve the model, a method based on Genetic Algorithm (GA) and Point wise Stationary Fluid Flow Approximation (PSFFA) was designed. Zehendner and Feillet (2013) quantified the benefits of using the truck appointment system for improving the service quality in CTs. A combined solution approach is adopted to solve the proposed mixed integer linear programming model. A DES validates the results obtained by the optimization model in a stochastic environment. The management of export container arrivals is studied by Chen and Yang (2010). They proposed an integer programming model and solved it using genetic algorithm (GA). In their paper, the transportation cost is reduced by adopting a time window management program.

Studying queuing behavior in container terminal received some interest. Kim (2009) proposed a non-linear integer programming model integrated with a stationary queuing model. Both waiting and operation cost are reduced. Another stationary time-dependent queuing model was introduced by Chen et al. (2011). The authors analyzed time-dependent truck queuing

processes with service time distributions at gates and yards of a port terminal. However, it is improper to use stationary queuing models to stochastically analyze a queuing system that is non-stationary in nature (Chen et al., 2011). An appointment system designed by a non-stationary queuing model was introduced by Chen et al. (2013a). The authors proposed two appointment systems: static and dynamic. GA was used to solve the optimization problem and simulation to compare results. Chen et al. (2013) proposed a method called ‘vessel dependent time windows (VDTWs)’ to alleviation of gate congestion. A hybrid algorithm using GA and simulated annealing was used to solve the optimization problem.

Researchers targeted the environmental objective for lowering the carbon dioxide emissions. In this context, Chen et al. (2013b) developed a bi-objective model to minimize both the truck waiting times and truck arrival pattern change. A GA based heuristics was used to solve the model and resulted in reduction of truck emissions using a small shift in truck arrivals.

Simulation was used in solving landside problems such as congestion, waiting, resources idling and emissions. Sharif et al. (2011) used agent-based-simulation to minimize congestion at seaport terminal gates by using the provided real-time gate congestion information and simple logic for estimating the expected truck wait time. Also, Veloqui et al. (2014) provided a DES model for truck arrival at the gate and yard. Various scenarios were simulated to reduce queues by using a commercial simulation software (FlexSim CT). A recent DES model to reduce empty truck trips by implementing a coordinated truck appointments was proposed by Schulte et al. (2015). Their model reduced the emissions but not the congestion.

Previous research also addressed yard operations like yard crane scheduling and container stacking. Online stacking rules are studied by Borgman et al. (2010). A DES tool is used to improve the yard efficiency. The impact of truck announcement on online stacking rules was studied by Asperen et al. (2011) as an extension work for Borgman et al. (2010). A DES model showed the benefits of using the truck announcement for increasing stacking yard efficiency. A new concept of chassis exchange terminal (CET) is presented by Dekker et al. (2013) to reduce terminals congestion by using simulation. Geith et al. (2014) provided an integer programming formulation for container pre-marshalling problem to minimize the containers mis-overlays with the minimum number of container movements. In a later work, Geith et al. (2016) used a variable chromosome length GA to solve the container pre-marshalling large size problems.

More studies were performed to improve the efficiency of yard operations using the information from gates.

Zhao and Goodchild (2010a) used simulation to evaluate the use of truck arrival information to reduce container repositioning during the import container retrieval processes. Zhao and Goodchild (2010b) also investigated the effectiveness of truck arrival information in reducing truck transaction times within container terminals by using the revised difference heuristic. A computer-based simulation is developed. Zhao and Goodchild (2013) extended their work and provided a hybrid approach of simulation and queuing theory to model the container retrieval operation and estimate the crane productivity and truck turn-time. The authors quantified the impact of using a truck appointment system on the yard efficiency of container terminals.

Smoothing truck arrivals in peak hours became a necessity for both container terminals and trucking companies. To achieve this goal, Phan and Kim (2015) addressed a negotiation process among multiple trucking companies and a terminal for smoothing truck arrivals in peak hours. A nonlinear mathematical model is formulated to develop an appointment system using the proposed negotiation process. They recommended to use simulation to validate their procedure of solution. The most recent paper by Li et al. (2016) discussed the deviation of trucks arrival from their appointments. DES is used to evaluate the performance of their proposed solution strategies.

PROBLEM DESCRIPTION AND SOLUTION FRAMEWORK

In container terminals, export container are brought to the seaport by external trucks and import containers are picked by external trucks to be delivered to the hinterland. One of the most imperative issues for the external trucks is the long truck turn time (TTT). The following equation describes the TTT:

$$TTT = T_{wg} + T_{sg} + T_{wy} + T_{sy} + T_{xg} \quad (1)$$

T_{wg} : waiting time at gate.

T_{sg} : service time at gate.

T_{wy} : waiting time at yard.

T_{sy} : service time at yard.

T_{xg} : time spent at gate exit.

Terminal operators need to reduce the TTT as much as possible. The truck turn time has direct and indirect impacts on the terminal efficiency. As direct impacts, shorter waiting times and service times reduce the congestion outside the gates and within the yard area. In addition, decreasing the turn time increases the terminal throughput and reduces the processes cost. Indirectly, emissions are reduced by less waiting and idling of the trucks and terminal equipment.

The gate operators usually force the trucks to wait outside the terminal or at specific waiting areas within the terminal to avoid the congestion at yard. This creates

new congestions at the gates. Moreover, not all terminals have enough waiting space within the terminals. The appointment systems for the external trucks are considered a managerial solution for the long TTT and congestions. There are many factors that affect the TTT like the gate capacity, gate working hours, resources within the terminal and truck arrival patterns. In this paper a discrete event simulation model to study the effect of the arrival patterns on the TTT and how the arrival patterns can be considered in improving the truck appointment system is presented. Also, an approach to develop an optimum appointment schedule is proposed based on the simulation results.

In the literature, many studies evaluated the impact of using an appointment system and arrival information to reduce the waiting, congestion, cost and emissions. Some of them tried to reduce the truck arrival to achieve these goals. Huynh and Walton (2008, 2011) studied the effect of limiting the truck arrivals to reduce (TTT). However, the reduction of truck arrivals disrupts the containers delivery times. To our best knowledge, studying the effect of arrival patterns of the external trucks was not considered enough in the previous literature.

THE PROPOSED SIMULATION MODEL AND EXPERIMENTS

Fig. 3 shows the 3D model for the main areas of the container terminal: quayside, yard area, and gate side. This simple model considers one gate, one yard block, one yard crane, one quay crane, and one vessel. FlexSim CT software is used as a special simulation software for container terminal operations. FlexSim CT provides the advantage of the built-in CT library. This library enables the modeler to save the time of building the container terminal objects and planners.

Both export and import container arrivals are simulated. Various truck arrival patterns are tested under stochastic situation. Investigating the truck turn time changes with various patterns is provided in the paper. The main target is to illustrate how to keep the TTT at minimum and maintain the gate throughput as high as possible. One vessel is proposed to reach the terminal at the beginning of the week. The vessel is assumed to deliver 200 import containers and to be charged with 200 export containers which come via the gate by the external trucks. Similarly, the imports will be picked up during the week by the external trucks. The model parameters are shown in Table 1. These parameters are the default parameters in FlexSim CT with little modification according to some practical experience. In Fig. 4 the process flow is described.

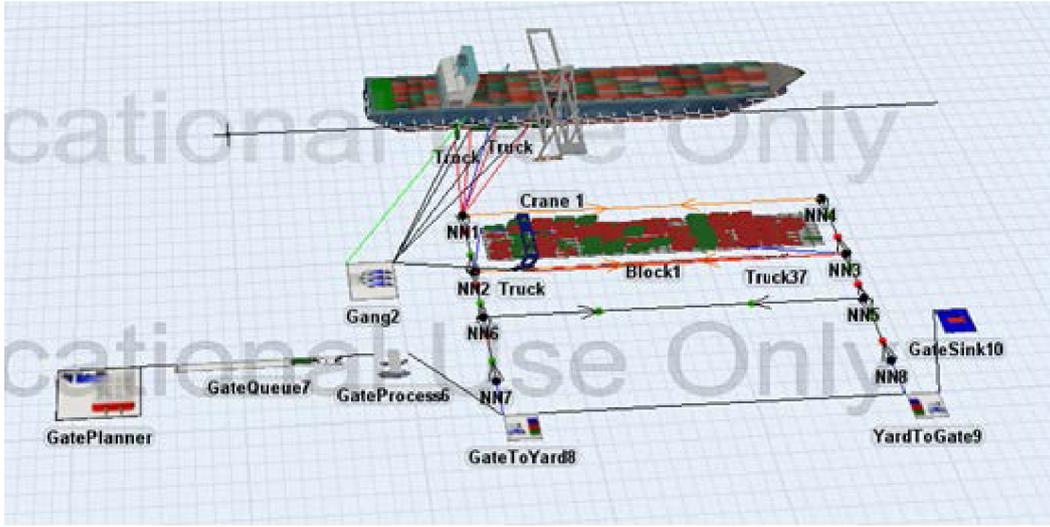


Fig. 3 Discrete Event Simulation Model

Table 1: Simulation Model parameters

Gate parameters	
Working hours/day	6:00 am – 8:00 pm
Trucking speed (max)	300 m/min
Process time	Triangular (5, 15, 10)
Pick up travel time	triangular(2,5,3)
Drop off travel time	triangular(2,5,3)
Dwell time variability	12 hrs.
Yard parameters	
Container dwell time	3 Days
Pick up time	triangular(0.2,2.0,0.5)
Drop off time	triangular(0.2,2.0,0.5)
Yard crane speed (max)	180 m/min
Quayside parameters	
Quay care speed (max)	120 m/min

Five arrival patterns are tested; the default arrival pattern (Def.), increasing arrival pattern (Inc.), decreasing arrival pattern (Dec.), uniform arrival pattern (Uni.), and distributed-peak arrival pattern (Dist.) (Fig. 5). The vertical axis represents the number of external trucks and the horizontal axis represents the day's working hours. These five patterns are developed using the "Gate Planner". The gray bars represent the arrival pattern that is needed to be matched by the gate planner. At the beginning of each week, red (dark) bars are drawn over the pattern to show the actual number of containers scheduled for each hour (both pickup and drop-off).

The default pattern shows that the arrivals reach a peak at the middle of the day. In some cases, the peaks the arrivals are at the end or beginning of the day. This situation is simulated using the increasing and decreasing patterns. The uniform arrival pattern proposes a stable level of arrival over the day. For the

distributed-peak arrival pattern, the arrival peaks are distributed to the beginning and the end of the working day. Fig. 5 shows a screen shot for the arrival patterns and the actual arrivals in a specific day after running the model.

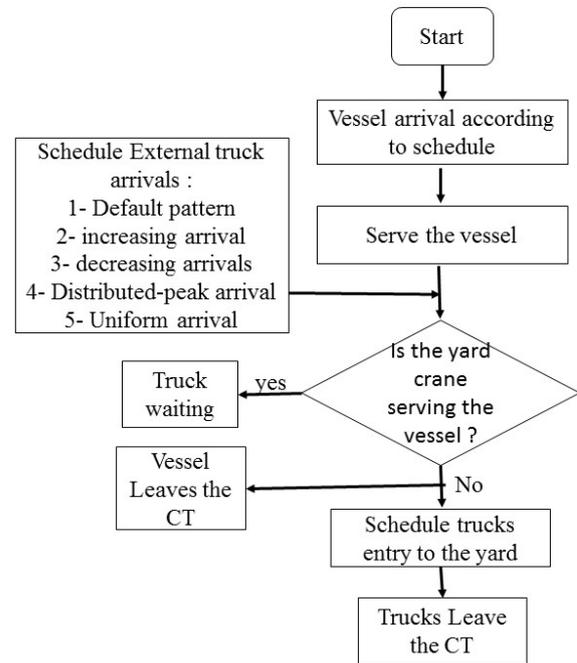


Fig. 4 Process Flow

RESULTS

The model is run for 50 replications in the steady state with a one week length for each replication, and the performance statistics are collected for each pattern as illustrated in (Table 2.).

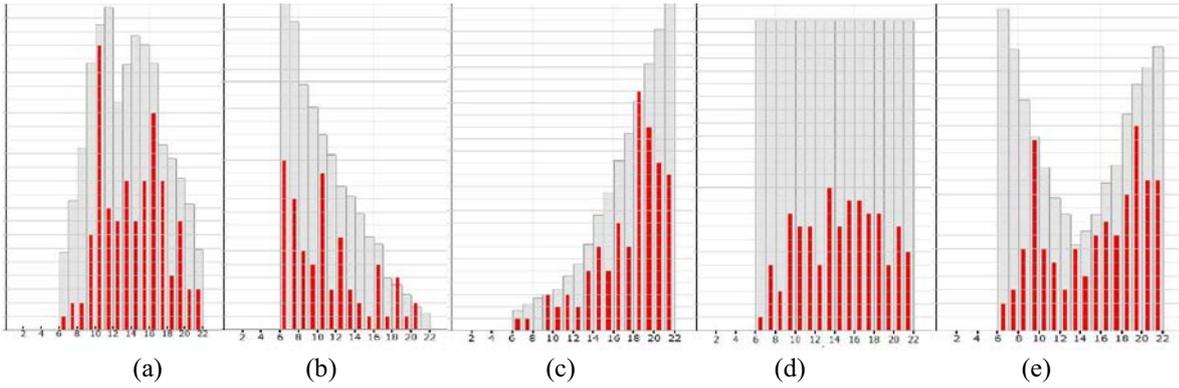


Fig. 5 Arrival Patterns: (a) Default Arrival, (b) Decreasing Arrival, (c) Increasing Arrival, (d) Uniform Arrival , and (e) Distributed-Peak Arrival

The performance metrics studied are the average TTT, minimum and maximum TTT, and maximum queue length at the gate. The average truck turn time exhibited different values for each pattern. Uniform arrivals exhibited the minimum value of TTT among all patterns. The decreasing arrival pattern had the maximum TTT of about 6 hrs. As noticed from the results, the average gate throughput is kept at the same level. This means that the arrival pattern can be used to reduce the TTT without limiting truck arrivals or reducing terminal productivity.

Table 2: Simulation Results

Performance metrics	Def.	Inc.	Dec.	Uni.	Dist.
Average TTT (min)	29.3	48.8	55.1	20.5	25.1
Min TTT (min)	11.4	11.4	11.3	11.4	11.3
Max TTT (min)	266.5	417.2	361.6	198.9	200.4
Max Queue length (trucks)	26	41	35	19	19
Average Gate throughput/week	398	398	398	398	398
Variance of gate throughput/week	80.4	71.3	69.2	71.6	67.7

From the results, the uniform arrival pattern exhibited the lowest average of TTT. A t-test with 0.05 significance level is conducted to compare the TTT averages for the uniform pattern and distributed- peak pattern. The t-test results showed that the average TTT of the uniform arrival is significantly less than the distributed-peak arrival's average. Based on this result, the simulation work is extended to obtain the best arrival

schedule per each hour. To do that, 10 replications for the uniform arrivals are performed and the scheduled arrivals for each hour of the working day is recorded. Fig. 6 shows the average numbers of trucks which are expected to achieve the minimum truck turn time and max gate throughput. Using truck schedules, the terminal operators can design an appointment system which avoids the congestions and long queues at the gates.

CONCLUSIONS AND FUTURE WORK

Truck delays are a common problem in container terminals. Long truck queues affect the efficiency of seaports and cause congestion problem at gates and yards. To reduce the turn time of external trucks, one needs to dig for factors that influence these delays. One of the most important factors is the arrival pattern of the trucks. In this paper a simple DES model is developed to show how the effect of the arrival patterns on the truck turn time and gate congestion can be studied. The results showed that shorter delays at CTs could be achieved without reducing number of truck arrivals or increasing the terminal resources. Moreover, simulation can help in designing the appointment systems in CTs.

For future work, we intend to study truck delays at Alexandria seaport in Egypt and introduce some strategies to obtain the optimum arrival patterns for truck arrivals. In addition, the appointment system shall consider the stochastic nature of inter-arrival times and other important factors.

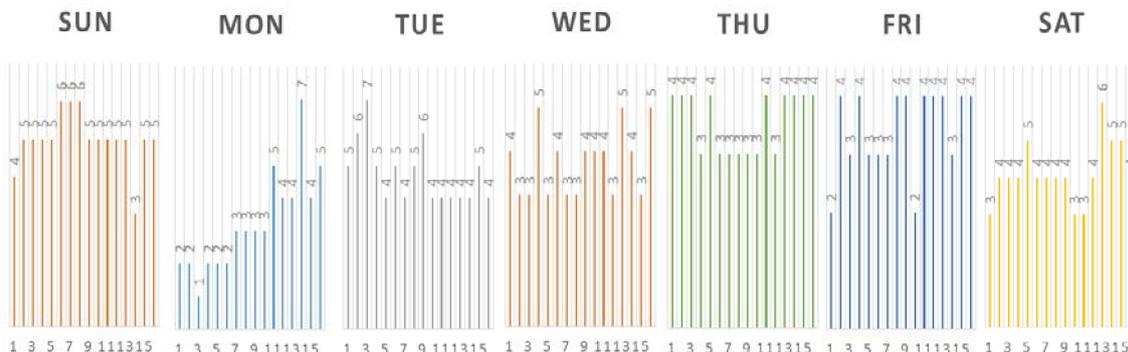


Fig. 6 Scheduled arrival during the week according to a uniform arrival pattern for 16 working hours/day

REFERENCES

- Angeloudis P. and Bell M. 2011. "A review of container terminal simulation models." *The flagship journal of international shipping and port research* VOL. 38, NO. 5, 523–540
- Asperen E., Borgman B. and Dekker R. 2011. "Evaluating impact of truck announcements on container stacking efficiency" *Flexible Services and Manufacturing Journal*, 25:543–556
- Carlo H. J., Vis I. and Roodbergen K. 2013. "Storage yard operations in container terminals: Literature overview, trends, and research directions" *European Journal of Operational Research*.
- Chen G. and Yanga Z. 2010. "Optimizing time windows for managing export container arrivals at Chinese container terminals." *Maritime Economics & Logistics* (2010).
- Chen G., Govindan K. and Golias M. 2013b. "Reducing truck emissions at container terminals in a low carbon economy: Proposal of a queueing-based bi-objective model for optimizing truck arrival pattern." *Transportation Research Part E*, 55 (2013) 3–22
- Chen G., Govindan K. and Yang Z. 2011. "Designing terminal appointment system with integer programming and non-stationary queueing model." *Technique Report, University of Southern Denmark*.
- Chen G., Govindan K. and Yang Z. 2013. "Managing truck arrivals with time windows to alleviate gate congestion at container terminals." *Int. J. Production Economics*, 141 (2013) 179–188
- Chen G., Govindan K., Yang Z. and Choi T. and Jiang L. 2013a. "Terminal appointment system design by non-stationary $M(t)/Ek/c(t)$ queueing model and genetic algorithm." *Int. J. Production Economics*, 146 (2013) 694–703
- Chen X., Zhou X. and List G. 2011. "Using time-varying tolls to optimize truck arrivals at ports." *Transportation Research Part E*, 47 (2011) 965–982
- Dekker R., Heide S., Asperen E. and Ypsilantis P. 2012. "A chassis exchange terminal to reduce truck congestion at container terminals." *Flexible Services and Manufacturing Journal* 25:528–542
- Gharehgozli A., Roy D. and Koster R. 2014. "Sea Container Terminals: New Technologies, OR models, and Emerging Research Areas." *ERIM Report Series reference number ERS-2014-009-LIS*.
- Gheith M., Eltawil A., Harraz N., Mizuno S., "An integer programming formulation and solution for the container pre-marshalling problem", *44th Int. Conf. on Computers and Industrial Engineering, Istanbul, Turkey*.
- Gheith M., Eltawil A., Harraz N., 2016 "Solving the container pre-marshalling problem using variable length genetic algorithms", *Engineering Optimization*, Vol 48, issue 4, pp 687-705
- Guan C. and Liu R. 2009. "Container terminal gate appointment system optimization." *Maritime Economics & Logistics* 11, 378–398. doi:10.1057/mel.2009.13
- Huynh N. 2009. "Reducing Truck Turn Times at Marine Terminals with Appointment Scheduling." *Transportation Research Record: Journal of the Transportation Research Board*.
- Huynh N. and Hutson N. 2008. "Mining the Sources of Delay for Dray Trucks at Container Terminals." *Transportation Research Record: Journal of the Transportation Research Board*.
- Huynh N. and Walton C. 2008. "Robust Scheduling of Truck Arrivals at Marine Container Terminals." *Journal of transportation engineering*.
- Huynh N. and Walton C. 2011. "Improving Efficiency of Drayage Operations at Seaport Container Terminals Through the Use of an Appointment System." *Handbook of Terminal Planning* Volume 49 of the series Operations Research/ Computer Science Interfaces Series pp 323-344
- Karam A., Eltawil A., Harraz N. 2014, "an improved solution for integrated berth allocation and quay crane assignment problem in container terminals", *44th Int. Conf. on Computers and Industrial Engineering, Istanbul, Turkey*.
- Kim S. 2009. "The toll plaza optimization problem: Design, operations, and strategies." *Transportation Research Part E* 45 (2009) 125–137
- Li N., Chen G., Govindan K. and Jin Z. 2016. "Disruption management for truck appointment system at a container terminal: A green initiative." *Transportation Research Part D*.
- Namboothiri R., Erera A. 2008. "Planning local container drayage operations given a port access appointment system." *Transportation Research Part E* 44 (2008) 185–202.
- Phan M. and Kim K. 2015. "Negotiating truck arrival times among trucking companies and a container terminal." *Transportation Research Part E* 75 132–144
- Schulte F., Gonzalez R. and Voß S. 2015. "Reducing Port-Related Truck Emissions: Coordinated Truck Appointments to Reduce Empty Truck Trips." *Computational Logistics* Volume 9335 of the series Lecture Notes in Computer Science pp 495-509
- Sharif O., Huynh N. and Vidal J. 2011. "Application of El Farol model for managing marine terminal gate congestion." *Research in Transportation Economics* 32 (2011) 81e89
- Stahlbock R. Voß S. 2008. "Operations research at container terminals: a literature update." *OR Spectrum* 30:1–52
- Steenken D., Voß S. and Stahlbock R. (2004) "Container terminal operation and operations research – a classification and literature review." *OR Spectrum* 26: 3–49
- Veloqui M., Turias I., Cerbán M., González M., Buiza G. and Beltrán J. 2014. "Simulating the landside congestion in a container terminal. The experience of the port of Naples (Italy)," *XI Congreso de Ingeniería del Transporte (CIT 2014)*.
- Zehendner E. and Feillet D. 2013. "Benefits of a truck appointment system on the service quality of inland transport modes at a multimodal container terminal." *European Journal of Operational Research* 235 (2014) 461–469
- Zhang X., Zeng O. and Chen W. 2013. "Optimization Model for Truck Appointment in Container Terminals." *13th COTA International Conference of Transportation Professionals*.
- Zhao W. and Goodchild A. 2010a. "The impact of truck arrival information on container terminal rehandling." *Transportation Research Part E* 46 (2010) 327–343 (2010a)
- Zhao W. and Goodchild A. 2010b. "Impact of Truck Arrival Information on System Efficiency at Container Terminals." *Transportation Research Record: Journal of the transportation Research Board*
- Zhao W. and Goodchild A. 2013. "Using the truck appointment system to improve yard efficiency in container terminals." *Maritime Economics & Logistics* 15.

3D Simulation Modeling of Yard Operation in a Container Terminal

Jingjing Yu
Chen Liang
Guolei Tang *

Faculty of Infrastructure Engineering
Dalian University of Technology
Dalian 116024, China

* E-mail: tangguolei@dlut.edu.cn

KEYWORDS

3D Simulation and Modeling, Container Terminal, Yard Operation.

ABSTRACT

The yard operation plays an important role in the daily running of a container terminal. Whether the efficiency of a terminal can be improved depends to a significant extent on the operation of its container yard. In order to analyze the yard operation, this paper builds a 3D simulation model to simulate and visualize the yard operation in a container terminal. First, we study the characteristics of yard layout and present the logical model of container yard operation with rubber-tired gantry cranes and trucks. Then, a 3D simulation model of yard operation is implemented, which includes model setup module, container terminal layout module, horizontal transport module, yard operation module, statistics analysis module, and 3D animation module. Finally, the implemented model is applied in practice to examine the impact of reshuffle operation, and the numbers of internal trucks and yard cranes on the efficiency of yard operation. And the results show the proposed simulation model performs well and is helpful for exploring yard operation effectively.

1. INTRODUCTION

Due to the boom in world trade, over 90% of cargo currently transported worldwide is shipped as containerized cargo (Liu et al. 2002). The container yard as a central part of cargo stacking and transport has a significant impact on the whole operation of a container terminal. Therefore, to design and operate a successful container terminal, an effective model is needed to help the planners to evaluate and explore the efficiency of yard operation considering the stochastic characteristics of port system.

Therefore, many researchers have applied computer simulation to study the operation of container terminals. Petering (2009) analyzed the effect of block width and storage yard layout on the performance of a container transshipment terminal. He simulated dozens of yard configurations to determine the optimal block width

(Petering 2009). Böse et al. (2000) compared different strategies of trucks dispatching to yard cranes to reduce the time in port for the vessels. Veeke and Ottjes (2002) used computer simulation to provide a decision support for the extension of Rotterdam Port. Gu et al. (2007) applied dynamic simulation in order to provide the operators and designers some advice on the plan and design of a container yard.

The objective of this paper is to establish a 3D simulation model of container yard operation. we focus on the implementation of traffic simulation of horizontal transport and 3D animation of yard operation, which are main contributions of this paper.

2. ANALYSIS OF YARD OPERATION

2.1 Layout of Container Yard

The container yard is composed of storage blocks and driving lanes separating those blocks (see Figure 1). The block structure is determined by equipment type used and, more importantly, by the options for transferring a container between a storage block and a horizontal means of transport. Therefore, the yard layout is defined by the organization of the driving lanes, by the number of driving lanes, by the orientation of the storage blocks, the block structure and the design of the storage blocks (Bish et al. 2001).

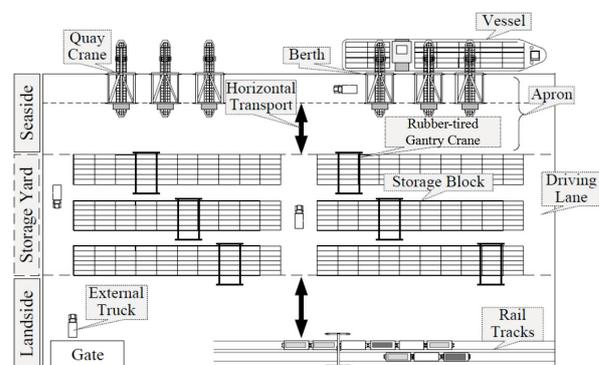


Figure 1: Schematic Structure of a Typical Container Terminal

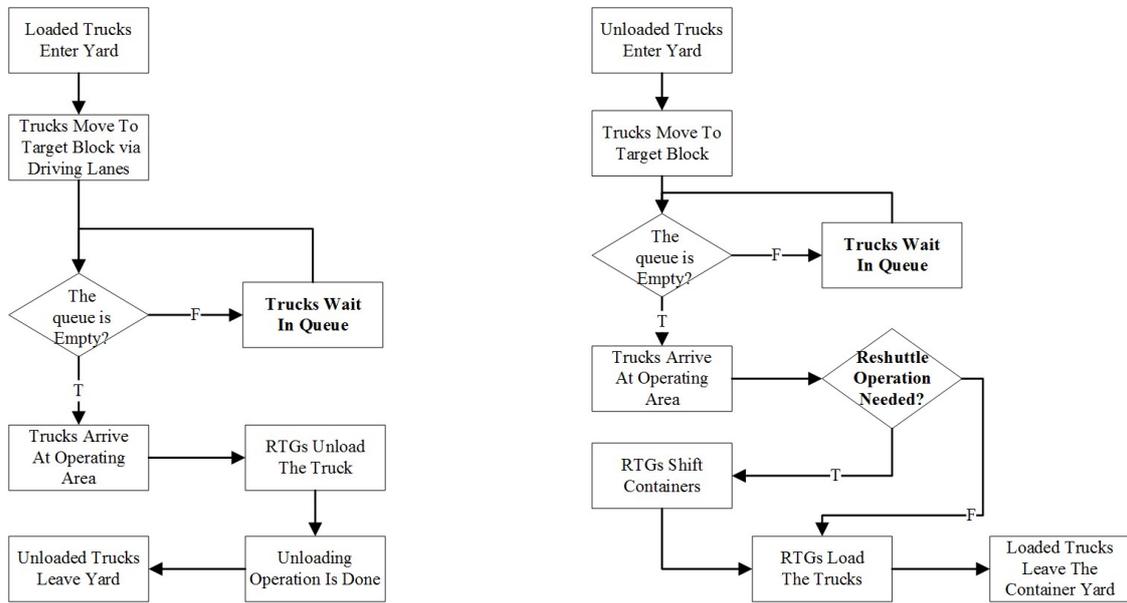
As shown in Figure 1, in most container terminals in China, the orientation of the storage blocks is parallel to the quay (in the case of a single straight quay wall). And rubber-tired gantry cranes (RTGs) are widely used for the yard operations (e.g., loading/unloading, stocking and reshuffle operation), and internal and external trucks for the horizon transport between quay and storage yard, as well as between storage yard and landside interfaces (Ji et al. 2010).

2.2 Logical Model of Container Yard Operation

Figure 2 shows the logic model of yard operations with RTGs and internal and external trucks. The trucks loaded containers travel through driving lanes into the yard, where the container is stacked into a block. As illustrated in Figure 2 (a), when the truck arrives at transfer lane of target block, the RTGs in this block

unload the container from the truck, and stack it into a block, if the queue of transfer lane is empty.

The container remains in the block until it is collected by another carrier (e.g., vessel, truck). In this case the container will be retrieved from the block and passed to an internal or external truck which conveys the container to its destination at the quay or the landside. However, a container might be requested from a stack in which other containers are stored upon the requested container. In this case the containers on top of the requested container have to be repositioned within the block, as shown in Figure 2 (b). The repositioning moves of containers within a block to retrieve another container are called reshuffle operation. Therefore, the main processes in the yard are thus the loading and unloading of container trucks, the stacking of containers and the retrieval of containers.



(a) Unloading and Stacking Operation (b) Loading and Retrieval Operation
Figure 2: The Logical Model of Container Yard Operation with RTGs and Trucks

3. 3D SIMULATION MODELING OF YARD OPERATION IN A CONTAINER TERMINAL

This paper implements a 3D simulation model of container yard operation by using AnyLogic simulation software (AnyLogic 2016). According to the characteristics of yard operation and the relationship with other terminal operations, the model consists of six sub-models, including model setup module, container terminal layout module, horizontal transport module, yard operation module, statistics analysis module, and 3D animation module.

3.1 Model Setup Sub-Model

In this sub-module, we define a series of variables and parameters (listed in Table 1), to describe the characteristics of entities and resources in the simulation

model. So it makes the simulation model flexible as it can be applied to different settings by properly tuning the several variables and parameters in the model.

Table 1: Variables/Parameters of Entities

Entities/Resources	Variables/Parameters
Rubber-Tired Crane	number, size, action
Quay Crane	number, size, action
Container	size, color, number
Internal Truck	size, speed, number
External Truck	arrivalPattern, speed
Storage Yard	type, size, height, number
Seaside	berth, length
Landside	parkingLots, location, number
Driving Lanes	nodes, lines, direction

3.2 Container Terminal Layout Sub-Model

This sub-model is used to visualize the layout of a specific container terminal, including the seaside, the storage yard (e.g., storage blocks and transfer lanes), the landside (parking lots) and driving lanes network.

3.3 Horizontal Transport Sub-Model

This sub-model is used for trucks scheduling, which chooses the optimal path according to the traffic of driving lanes between the apron and storage yard as well as between storage yard and landside interfaces.

3.4 Yard Operation Sub-Model

(1) Queue of Trucks in Storage Blocks

The queue is used to store the trucks waiting for a RTG for loading/unloading containers.

(2) Unloading Operation Module

When a truck with containers arrives at the target transfer lane, it first seizes an idle RTG by “Seize” module in AnyLogic, which unloads the container from the truck, and steps to “Stacking Operation Module”.

(3) Stacking Operation Module

The seized RTG stacks the container into the specified block according to the attribute of container entity. Once the container is stacked, the RTG is released via “Release” module in AnyLogic, and this model updates the states of storage blocks, and the location of the container. Finally, the RTG goes back to the original location and serves the next truck.

(4) Retrieval Operation Module

When a truck arrives at the specific transfer lane to collect a container, it first seizes an idle RTG by “Seize” module in AnyLogic. Then the RTG determines whether the reshuffle operation is needed according to the location of the target container. If the reshuffles are not needed, the RTG retrieves the target container. Otherwise, the RTG moves containers within a block to retrieve the target container, according to implemented function. The function is used to determine the number and locations of containers that need reshuffles, and the sequence of reshuffles moves. Finally, the RTG steps “Loading Operation Module”.

(5) Loading Operation Module

The RTG loads the target container on the truck, and is released via “Release” module in AnyLogic. Once the container is loaded, the RTG goes back to the original location and serves the next truck. And this module updates the state of storage blocks synchronously.

3.5 Statistics Analysis Sub-Model

This sub-model records the real-time information, including the locations of the target containers, the number and locations of containers in each storage block, etc. And the values of indicators for yard

operation performance are also analyzed, such as the utilization ratio of yard cranes.

3.6 3D Animation Sub-Model

This sub-model is used to realize a 3D animation of yard operation, which helps the planners to intuitively identify the bottleneck that may be encountered during the operation of a container terminal.

(1) Horizontal Transport Animation

Figure 3 shows the 3D animation of horizontal transport traffic of driving lanes between storage yards and the seaside or gates. In this way, it is easy for the planner to check traffic congestion that might occur in the container terminal operation. Note that, in this animation, the traffic simulation is also included, so the simulated times of horizontal transport are more accurate as the effect of traffic congestion in transfer lanes is considered. Therefore, traffic simulation implementation is one contribution of this paper.

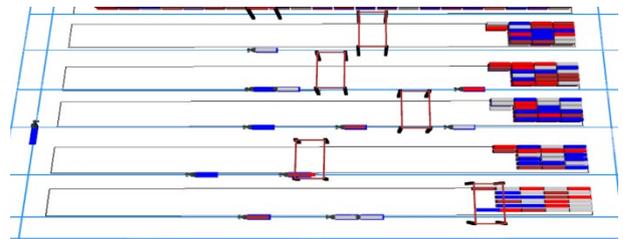


Figure 3: Horizontal Transport Animation

(2) Yard Operation Animation

The 3D animations of yard operation including loading/unloading operation (seen in Figure 4 and 5), stacking and retrieval operation (seen in Figure 6 and 7), are implemented according to the processes discussed in the Section 3.4.

In the yard operation animation, a series of operations, including lifting, translation and dropping are accomplished in order with the speeds set by the model setup sub-model, which is another contribution of this paper.

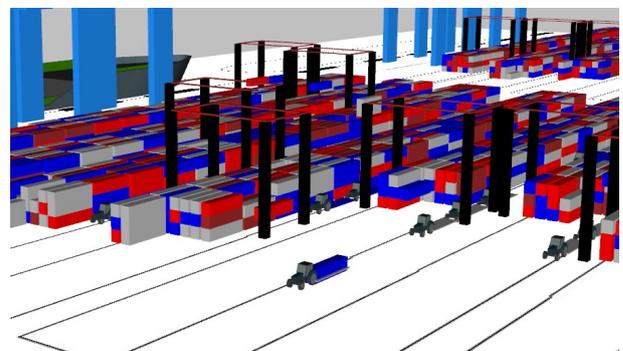


Figure 4: Loading Operation Animation

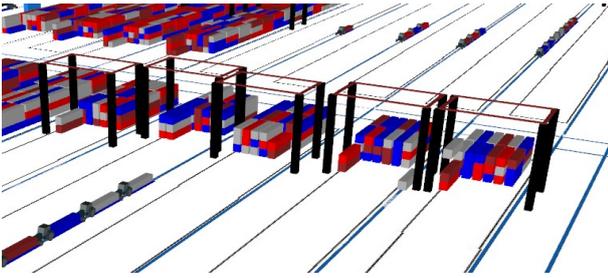


Figure 5: Unloading Operation Animation

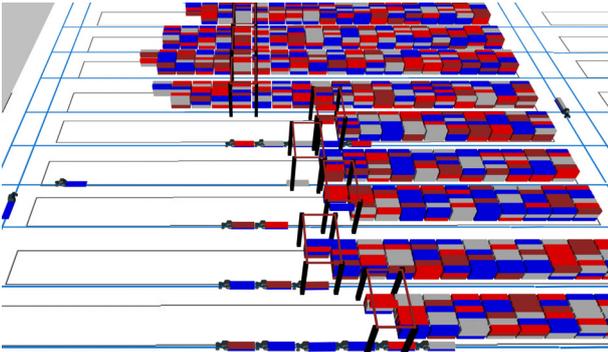


Figure 6: Stacking Operation Animation

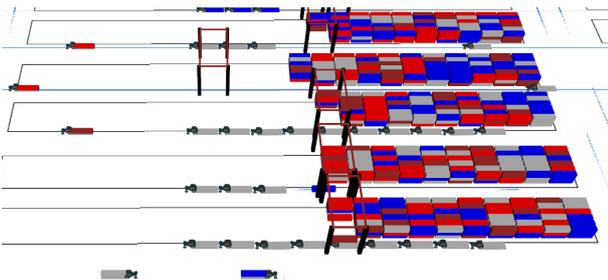


Figure 7: Retrieval Operation Animation

(3) 3D Model of Entities and Resources

The containers established in the model are standard 40-foot containers, and the external dimensions are 2.9m (height) * 2.44m (width) * 12.2m (length). The trucks in the model can transport a 40-foot container, and the length of trailers is 12.5m. The yard cranes are RTGs of SRTG5223S, with the dimensions of 26.5m long, 26.5m high and 15.2m wide, and the maximum lifting height of 18.2m.

3.7 Model Verification and Validation

This model is verified to confirm that it is correctly implemented with respect to the process of yard operation. First, we implement the model through sub-models and each sub-model is individually examined by a subject-matter expert. Secondly, we use tracing approach in AnyLogic throughout the development of simulation model to check if the logic implemented in the model is as intended. Finally, we use 3D animation to observe traffic flows, operation processes and the queues in yard, and verify simulation model logically.

The average time of trucks from the apron to the locations for loading/unloading and the average time of quay cranes to operate on single container are used to verify the simulation model. Therefore, we compare the simulation results with manual calculations as shown in Table 2. Since the discrepancies between the manual calculations and the simulation results are very small, the simulation model proposed in this paper can be used to reflect the impact of reshuffle operation, and the numbers of internal trucks and yard cranes on the efficiency of yard operation.

Table 2 Comparison between Manual Calculations and Simulation Results

Items	Average time of trucks (min)	Average time of gantry cranes(min)
Manual calculations	1.516	4.56
Simulation results	1.515	4.57
Errors (%)	0.065	0.219

4. CASE STUDY

The case study considers a container terminal with 2 70000-DWT berths. As shown in Figure 8, the container yard includes four rows of export blocks and five rows of import blocks. The rubber-tired gantry cranes are used for the yard operations, and internal and external trucks for the horizon transport between storage yard and seaside and gate. To explore the impact of reshuffle operation, and the numbers of internal trucks and yard cranes on the efficiency of yard operation, we implement a 3D simulation model to simulate the yard operation in this terminal.

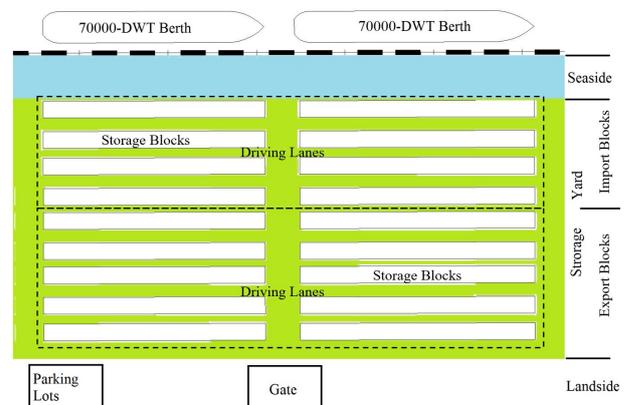


Figure 8: Schematic Structure of the Container Yard

4.1 Model Parameters

According to MTPRC (Ministry of Transport of the People’s Republic of China, 2014), the empty trucks and the loaded trucks move at speeds of 35km/h and 25km/h respectively. And the containers loaded and unloaded of a 70000-DWT container ship range from 2000 TEUs to 2250 TEUs. The specifications of container, trucks and RTGs refer to the Section of “3D Amination Sub-Model”.

4.2 Simulation Experiments

We change some parameters (e.g., the number of internal trucks serving a quay crane, the number of yard cranes in the yard, and whether reshuffle operation is

considered), and establish five simulation experiments to explore their impacts as shown in the Table 3.

4.3 Analysis and Discussion

The simulation model for this terminal has been verified and validated logically as described in Section 3.7 “Model Verification and Validation”. For each scenario, similar simulations are performed 10 times for a period of a week. Then the average utilization ratio of yard cranes and the average number of vessels served in a week can be obtained by running the simulation model as shown in the Table 4.

Table 3: The Parameters Setting of Simulation Experiments

No.	The Number of Quay Cranes Per Berth	The Number of Trucks Per Quay Crane	The Number of Yard Cranes Per Berth	Whether Reshuffle Operation is considered
1	4	5	9	YES
2	4	5	9	NO
3	4	5	18	YES
4	4	4	9	YES
5	4	6	9	YES

Table 4: Simulation Results of Different Simulation Experiments

No.	The Utilization Ratio of Yard Crane (%)	The Number of Vessels Served
1	72.3	8
2	70.0	11
3	62.0	14
4	64.7	8
5	71.0	9

(1) The Impact of the Reshuffle Operation on the Efficiency of Yard Operation

Based on the simulation results of Experiments 1 and 2, we can know that reshuffle operation can result in the lower utilization of yard cranes due to the longer occupied time. So that the vessels will spend more time in the port, and the number of vessels served at unit time is less, then it will lead to less competitiveness of the container terminal. Therefore, these reshuffle moves should be avoided as far as possible, as they slow down the retrieval of the requested container.

(2) The Impact of the Number of Trucks Per Quay Crane on the Efficiency of Yard Operation

Based on the simulation results of Experiments 1, 4 and 5, we conclude that with more trucks, the utilization ratios of yard cranes and the number of ships serviced

are stable. Therefore, more trucks per quay crane can partly improve the efficiency of yard operation.

(3) The Impact of the Number of Yard Cranes on the Efficiency of Yard Operation

Based on the simulation results of Experiments 1 and 3, we can know that when we increase the number of gantry cranes, the average utilization will decrease and the number of vessels that have finished loading and discharging process will increase. Therefore, more gantry cranes can shorten the queue of trucks and improve the efficiency of yard operation.

Therefore, the model can be used to investigate the impact of reshuffle operation, and the numbers of internal trucks and gantry cranes on the efficiency of yard operation.

5. CONCLUSIONS

The main contribution of this paper is to provide a 3D simulation model of container yard operation. And the proposed simulation model can simulate the real situation of container yard operation by controlling the exact coordinates of various processes, including stacking, moving, as well as lifting and dropping, etc. Based on the 3D simulation model, a good reference for operation planning of a container yard can be provided.

ACKNOWLEDGEMENTS

This paper is a partial result of the project supported by the National Natural Science Foundation, China (No. 51579035 & 51309049).

REFERENCES

- AnyLogic: AnyLogic. <http://www.anylogic.com/>. (Accessed from April 2016)
- Bish, E.K.; Leong, T.; Li, C.; Ng, W.C.J.; and Simchi-levi, D. 2001. "Analysis of a New Vehicle Scheduling and Location Problem". *Naval Research Logistic* 48, No.5, 363-385.
- Böse, J.; Reiners, T.; Steenken, D.; and Voß, S. 2000. "Vehicle Dispatching at Seaport Container Terminals Using Evolutionary Algorithms". In *Proceedings of the 33rd Hawaii International Conference on System Sciences-2000*.
- Gu, Y.; Dong, M.; and Liu, J. 2007. "The Simulation Modeling and Its Application in the Design of a Container Yard". *Journal of Wuhan University of Technology*. Department of Transportation Science & Engineering, Vol.31, No.4 (Aug), 633-636.
- Ji, S.; Wei, Z.; and Chen, J. 2010. *The Transportation and Equipment of Containers*. China Material Publication.
- Liu, C.; Jula, H.; and Ioannou, P.A. 2002. "Design, Simulation, and Evaluation of Automated Container Terminals". In *Proceedings of IEEE Transactions on intelligent transportation* (Mar). Vol 3, No.1, 12-26.
- Ministry of Transport of the People's Republic of China (MTPRC). (2014) Design code of general layout of sea ports, JTS 165-2013. Beijing: China Communication Press.
- Petering, M.E.H. 2009. "Effect of Block Width and Storage Yard Layout on Marine Container Terminal Performance". *Transportation Research Part E* 45, 591-610.
- Veeke, H.P.M and Ottjes, J.A. 2002. "A Generic Simulation Model for Systems of Container Terminals". *ESM*. 2002, 581-587.

AUTHOR BIOGRAPHIES



JINGJING YUN was born in Chaoyang City, Liaoning Province, China, and went to Dalian University of Technology, where she majored in port and waterway

engineering and obtained the Bachelor's Degree in 2015. Now, she is studying for a Doctor's Degree in the field of simulation for port and waterway engineering analysis. Her e-mail address is: yigaoyujingjing@mail.dlut.edu.cn, and her Webpage can be found at <http://port.dlut.edu.cn>.



CHEN LIANG was born in Tangshan City, Hebei Province, China, and went to Dalian University of Technology to study port and waterway engineering. In 2015, he obtained the Bachelor's Degree. Now, he is studying for a Master's Degree in the field of simulation for port and waterway engineering analysis. His e-mail address is: liangchen@mail.dlut.edu.cn, and his Webpage can be found at <http://port.dlut.edu.cn>.



GUOLEI TANG (Corresponding author) was born in Yantai City, Shandong Province, China, and went to Dalian University of Technology, where he obtained the Doctor's Degree in Hydrology and Water Resources in 2009. He worked for a couple of years for the simulation modeling in engineering, and now, he is leading a large research group in the field of simulation for port and waterway engineering analysis. His e-mail address is: tanguolei@dlut.edu.cn, and his Webpage can be found at <http://port.dlut.edu.cn>.

GENERIC REACTION-DIFFUSION MODEL FOR TRANSMISSION OF MOSQUITO-BORNE DISEASES: RESULTS OF SIMULATION WITH ACTUAL CASES

Cynthia Mui Lian Kon^{1,2} and Jane Labadin¹

¹Department of Computational Science and Mathematics
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
93400 Kota Samarahan Sarawak Malaysia

²Faculty of Engineering, Computing and Science
Swinburne University of Technology Sarawak Campus
Jalan Simpang Tiga
93350 Kuching Sarawak Malaysia
Emails: cynkonml@hotmail.com and ljane@unimas.my

KEYWORDS

malaria, diffusion, generic model, mosquito-borne diseases, transmission

ABSTRACT

Mosquitoes can cause a lot of suffering to humans by transferring diseases. Malaria is a mosquito-borne disease caused by the parasite *Plasmodium*. It is an acute public health issue in many countries and can be fatal. Considering the similarity in the transmission of mosquito-borne diseases, a generic spatial-temporal model for transmission of multiple mosquito-borne diseases was formulated. The main concern here is whether the numerical results produced by this reaction-diffusion generic model are comparable with actual cases. Here, the actual notified weekly cases for 36 weeks, which is from week 39 in 2012 to week 22 in 2013, for four districts in Sarawak, Malaysia, namely Kapit, Song, Belaga and Marudi are compared with simulations of the generic model. The random movement of human and mosquito populations are taken into account. It is discovered that the numerical results are in good agreement to the actual malaria cases in the four districts.

INTRODUCTION

Mosquito-borne diseases such as dengue, yellow fever, filariasis, Japanese Encephalitis, Zika fever and malaria are a few diseases which can be transmitted from human to human through mosquito bites. A mosquito will ingest a pathogen when it takes a blood meal on an infectious human. Next, when it bites a susceptible human, it injects saliva and anti-coagulants

into the human's blood. According to the World Health Organization, there were estimated 214 million malaria cases reported worldwide and 438000 deaths in 2015 (World Health Organization, 2015).

Disease models are essential in the armory of epidemiological devices as they can be utilized in spite of limitations in data. The mathematical models are based on the understanding of the dynamics of the host and pathogen and are used to provide forecast of the prevalence of an infection. The factors which contribute to the spread of the infection can be identified and to determine the best control measures to eradicate them. These models use mathematical equations to interpret the dynamics of diseases. Henceforth, experiments can be done without the necessity to carry it out in real life which might be unethical. By doing so, questions such as what-if can be answered. Deterministic models consist of ordinary differential equations, for example models by Ross (1910), Anderson and May (1979), Tumwiine et al. (2007), and Labadin et al. (2009). The compartmental mathematical models illustrate whether the disease will prevail or dies out in a population in time.

Researchers have been examining the impact of spatial heterogeneity and movements on the spread and persistence of diseases. Consequences of control measures such as vaccination in local region and its outcome can be analyzed. There are a few approaches to modeling spatial spread. Metapopulation models (Cliff, 1992, Ding et al. 2012) divide a population to multiple discrete groups. There are two ways these models are constructed, the mobility approach and the cross-coupled metapopulation approach. Most metapopulation models are the cross-coupled approach which consider transmission within and between groups. Spatially continuous models such as reaction-diffusion models assume the population is distributed continuously across the environment and not in discrete

populations. Spatially continuous models (Kon and Labadin, 2005, Anit̃a and Capasso, 2012, Maidana and Yang, 2007) allow mathematical analysis of the general patterns of the spread of diseases. Lattice-based models group a particular grid site as a subpopulation and in cellular automata, iterations are assumed to occur only between neighbouring grids (Gibson, 1997). Geographic information systems (GIS) are used to secure, collect spatial data and when needed, regain and to depict the spatial facts. Hence, it can be utilized for data on populations, disease prevalence, environmental data and create a connection among them. Examples of application of GIS are carried out by Chaput et al. (2002) and Kitron (2000).

A reaction-diffusion generic model for mosquito-borne diseases was formulated (Kon and Labadin, 2015) based on the similarity in the manner of transmission of these infections. All mosquito-borne diseases are spread through vector mosquito; thus, this commonality is taken into account. A general model which incorporates both spatial and temporal factors as well as caters to the many different mosquito-borne diseases is profitable as most models constructed are for a specific disease. In this paper, the proposed new generic model will be discussed in the next section. After that, the reported weekly malaria cases in four districts in Sarawak are compared to simulations from the generic model for the spread of mosquito-borne infection. In the last section, conclusions and plans for future work are discussed.

MODEL FORMULATION

Mosquito-borne and vector-borne diseases compartmental models were deliberated and the similarities identified before formulating the generic model (Kon and Labadin, 2013). This is because the objective is to formulate a generic model which can be applied to different mosquito-borne diseases. The matching compartments used for vector-borne diseases are found to be susceptible and infectious for both human and vector population. Thus, the generic mosquito-borne diseases model consists of Susceptible-Infectious (SI) compartments for both human host and vector mosquito (Kon and Labadin, 2015). As we want to investigate the effect of spatial heterogeneity and movement of human and mosquito population on the spread of diseases, spatial factors are incorporated in the generic model. The common terms used for spatial spread were determined by studying the spatio-temporal disease models and they are found to be diffusion coefficients and location dependent parameters. Terms such as birth rate, death rate, force of infection and recovery rate were regularly included in vector-borne disease models. The focal point here is in the way these diseases are transmitted. The random movement of human and mosquito populations are included and they are depicted as random walks, where a group of dispersing humans and mosquitoes behave comparatively to a group of particles diffusing in

Brownian motion at large spatial scale (Cantrell and Cosner, 2004).

Here, the human population is divided into two compartments namely susceptible S_H and infectious I_H . The density of susceptible and infectious human populations are $S_H(t, x)$ and $I_H(t, x)$ where location x is considered. It is assumed that dynamics of total human population obey

$$\frac{\partial N_H(t, x)}{\partial t} = D_H \frac{\partial^2 N_H(t, x)}{\partial x^2} + \gamma + d_H N_H(t, x) \quad \text{where total}$$

human density is $N_H = S_H + I_H$. $D_H \frac{\partial^2 N_H(t, x)}{\partial t^2}$

represents the random movement of total human population across the environment. The diffusion coefficient D_H portrays the change in the rate of change of human movement. It is assumed that diffusion coefficient for both susceptible and infectious is constant. The mosquito population is also divided into susceptible S_M and infectious I_M . $S_M(t, x)$ and $I_M(t, x)$ are the spatial density of susceptible and infectious mosquito respectively. Moreover, total mosquito density is $N_M = S_M + I_M$, giving us the total human and mosquito density at any point x and time t are $N_H(t, x)$ and $N_M(t, x)$ respectively. An infectious mosquito transfers the infection if it takes blood meal on a susceptible human. The rate that susceptible human gets infected is $c \frac{g N_H(t, x)}{1 + g \lambda N_H(t, x)} \frac{I_M(t, x)}{N_H(t, x)}$ where c is the probability of transmission per bite from an infectious mosquito to a susceptible human, $\frac{g N_H(t, x)}{1 + g \lambda N_H(t, x)}$ is the

mean rate of bites per mosquito, and $\frac{I_M(t, x)}{N_H(t, x)}$ is the

probability an infectious mosquito bites a human. The biting rate is density-dependent on the total human population as done by Wang and Zhao (2011). In addition, the number of new infectious cases is

$$c \beta \frac{S_H(t - \tau_H, x)}{H(x)} I_M(t - \tau_H, x) \quad \text{with the latent period for}$$

human given as τ_H . Hence, the disease is considered to be transferred to an infectious human before the latent period. A susceptible mosquito gets infected when it takes a blood meal on an infectious human. The rate that susceptible mosquito gets infected is

$$b \frac{g N_H(t, x)}{1 + g \lambda N_H(t, x)} \frac{I_H(t, x)}{N_H(t, x)} \quad \text{where } b \text{ is transmission per}$$

bite from an infectious human to a susceptible mosquito and $\frac{I_H(t, x)}{N_H(t, x)}$ is the probability that a mosquito bites an infectious human, I_H . The number of new mosquito

infectious cases is $b\beta \frac{S_M(t-\tau_M, x)}{H(x)} I_H(t-\tau_M, x)$,

taking into account that contact occurred before the mosquito incubation period, τ_M .

The reaction-diffusion generic model for transmission of mosquito-borne infection is below:

$$\frac{\partial S_H(t, x)}{\partial t} = D_H \frac{\partial^2 S_H(t, x)}{\partial x^2} + \gamma - c \frac{g N_H(t, x)}{1 + g \lambda N_H(t, x)} \frac{I_M(t, x)}{N_H(t, x)} S_H(t, x) + r I_H - d_H S_H(t, x) \quad (1)$$

$$\frac{\partial I_H(t, x)}{\partial t} = D_H \frac{\partial^2 I_H(t, x)}{\partial x^2} + c \frac{g N_H(t-\tau_H, x)}{1 + g \lambda N_H(t-\tau_H, x)} \frac{I_M(t-\tau_H, x)}{N_H(t-\tau_H, x)} S_H(t-\tau_H, x) - (d_H + r) I_H(t, x) \quad (2)$$

$$\frac{\partial S_M(t, x)}{\partial t} = D_M \frac{\partial^2 S_M(t, x)}{\partial x^2} + \Lambda - b \frac{g N_H(t-\tau_M, x)}{1 + g \lambda N_H(t-\tau_M, x)} \frac{I_H(t-\tau_M, x)}{N_H(t-\tau_M, x)} S_M(t, x) - d_M S_M(t, x) \quad (3)$$

$$\frac{\partial I_M(t, x)}{\partial t} = D_M \frac{\partial^2 I_M(t, x)}{\partial x^2} + b \frac{g N_H(t-\tau_M, x)}{1 + g \lambda N_H(t-\tau_M, x)} \frac{I_H(t-\tau_M, x)}{N_H(t-\tau_M, x)} S_M(t-\tau_M, x) - d_M I_M(t, x) \quad (4)$$

Parameters used in the model above are as stated in Table 1. All parameters are assumed to be non-negative.

Table 1: Parameters used in the partial differential equations model

Parameters
diffusion rate for humans, D_H
diffusion rate for mosquitoes, D_M
human recruitment rate, γ
transmission probability per bite from $I_M \rightarrow S_H$, c
human death rate, d_H
recovery rate, r
mosquito recruitment rate, Λ
transmission probability per bite from $I_H \rightarrow S_M$, b
mosquito death rate, d_M
incubation period in humans, τ_H
incubation period in mosquitoes, τ_M

searching rate of a mosquito, g
time for a mosquito to consume blood per bite, λ

MODEL ANALYSIS

Since the generic mosquito-borne diseases model is for multiple mosquito-borne diseases, we want to investigate whether this model is able to reproduce malaria cases in four different districts in Sarawak. The actual cases are taken from Sarawak Weekly Epid News by Sarawak State Health Department as all malaria cases should be notified within 24 hours as reported by the Vector Borne Disease Control Section (Sarawak Health Department, 2012). Simulation results from the generic model are compared with actual prevalence in four districts in Sarawak, ie. Kapit, Song, Belaga and Marudi from week 39 in 2012 to week 22 in 2013, that is for 36 weeks. Parameters used are stated in Table 2. As this system is made up of nonlinear partial differential equations, the model is discretized using the finite difference method. Crank Nicolson method is used as it is unconditionally stable for diffusion equations (Thomas, 1995). Firstly, let us start with

equation (1) by writing it at the point $(x_i, t^{j+\frac{1}{2}})$. Thus

$$\frac{\partial S_H(x_i, t^{j+\frac{1}{2}})}{\partial t} \approx \frac{S_H(x_i, t^{j+1}) - S_H(x_i, t^j)}{\Delta t} \quad \text{will be written as } \frac{S_{H_i}^{j+1} - S_{H_i}^j}{\Delta t}$$

approximation for S_H at $(x_i, t^{j+\frac{1}{2}})$. The term

$$\frac{\partial^2 S_H(x_i, t^{j+\frac{1}{2}})}{\partial t^2}$$

is approximated using the average of second centered differences for $\frac{\partial^2 S_H(x_i, t^{j+1})}{\partial t^2}$ and

$$\frac{\partial^2 S_H(x_i, t^j)}{\partial t^2}$$

Similar approximation is carried out on the other populations, that is for infectious human, susceptible mosquito and infectious mosquito.

The equations after discretization and rearrangement are:

$$\begin{aligned} & S_{H_{i+1}}^{j+1} \left(\frac{-D_H}{2(\Delta x)^2} \right) + S_{H_i}^{j+1} \left(\frac{1}{\Delta t} + \frac{D_H}{(\Delta x)^2} + \frac{c\beta}{4H} (I_{M_i}^j + I_{M_i}^{j+1}) + \frac{d_H}{2} \right) \\ & + S_{H_{i-1}}^{j+1} \left(\frac{-D_H}{2(\Delta x)^2} \right) = S_{H_{i+1}}^j \left(\frac{D_H}{2(\Delta x)^2} \right) + \gamma + \\ & S_{H_i}^j \left(\frac{1}{\Delta t} - \frac{D_H}{(\Delta x)^2} - \frac{c\beta}{4H} (I_{M_i}^j + I_{M_i}^{j+1}) - \frac{d_H}{2} \right) + S_{H_{i-1}}^j \left(\frac{D_H}{2(\Delta x)^2} \right) \\ & + r \left(\frac{I_{H_i}^j + I_{H_i}^{j+1}}{2} \right), \end{aligned} \quad (5)$$

Table 2: Values of parameters used for Kapit, Song, Marudi and Belaga districts

Parameters	Kapit (dimensions)	Song	Marudi	Belaga	Source
D_H	10 km ² week ⁻¹	10	10	10	
D_M	10 km ² week ⁻¹	10	10	10	
γ	201 week ⁻¹	201	201	201	Department of Statistics Malaysia Sarawak (2013)
λ	3.5 week	3.5	3.5	3.5	Wang and Zhao (2011)
g	2.8 week ⁻¹	2.8	2.8	2.8	Chitnis (2005)
c	0.012	0.012	0.012	0.012	Nedelman (1984)
d_H	0.0000827 week ⁻¹	0.0000827	0.0000827	0.0000827	Department of Statistics Malaysia Sarawak (2013)
r	0.014	0.014	0.014	0.014	Williams and Boland (2002)
Λ	1056	1056	1056	1056	

b	0.47	0.47	0.47	0.47	Mehlhorn (2001)
d_M	0.0485 week ⁻¹	0.0485	0.0485	0.0485	
τ_H	1.4 week	1.4	1.4	1.4	Anderson and May (2001)
τ_M	1.4 week	1.4	1.4	1.4	Cox (2002)

$$I_{H_{i+1}}^{j+1} \left(\frac{-D_H}{2(\Delta x)^2} \right) + I_{H_i}^{j+1} \left(\frac{1}{\Delta t} + \frac{D_H}{(\Delta x)^2} + \frac{r}{2} + \frac{d_H}{2} \right) + I_{H_{i-1}}^{j+1} \left(\frac{-D_H}{2(\Delta x)^2} \right) = I_{H_{i+1}}^j \left(\frac{D_H}{2(\Delta x)^2} \right) + I_{H_i}^j \left(\frac{1}{\Delta t} - \frac{D_H}{(\Delta x)^2} - \frac{r}{2} - \frac{d_H}{2} \right) + I_{H_{i-1}}^j \left(\frac{D_H}{2(\Delta x)^2} \right) + \frac{c\beta}{H} \left(\frac{S_{H_i}^j + S_{H_i}^{j+1}}{2} \right) \left(\frac{I_{M_i}^j + I_{M_i}^{j+1}}{2} \right) \quad (6)$$

$$S_{M_{i+1}}^{j+1} \left(\frac{-D_M}{2(\Delta x)^2} \right) + S_{M_i}^{j+1} \left(\frac{1}{\Delta t} + \frac{D_M}{(\Delta x)^2} + \frac{b\beta}{4H} (I_{H_i}^j + I_{H_i}^{j+1}) + \frac{d_M}{2} \right) + S_{M_{i-1}}^{j+1} \left(\frac{-D_M}{2(\Delta x)^2} \right) = S_{M_{i+1}}^j \left(\frac{D_M}{2(\Delta x)^2} \right) + \Lambda + S_{M_i}^j \left(\frac{1}{\Delta t} - \frac{D_M}{(\Delta x)^2} - \frac{b\beta}{4H} (I_{H_i}^j + I_{H_i}^{j+1}) - \frac{d_M}{2} \right) + S_{M_{i-1}}^j \left(\frac{D_M}{2(\Delta x)^2} \right) \quad (7)$$

and

$$I_{M_{i+1}}^{j+1} \left(\frac{-D_M}{2(\Delta x)^2} \right) + I_{M_i}^{j+1} \left(\frac{1}{\Delta t} + \frac{D_M}{(\Delta x)^2} + \frac{d_M}{2} \right) + I_{M_{i-1}}^{j+1} \left(\frac{-D_M}{2(\Delta x)^2} \right) = I_{M_{i+1}}^j \left(\frac{D_M}{2(\Delta x)^2} \right) + I_{M_i}^j \left(\frac{1}{\Delta t} - \frac{D_M}{(\Delta x)^2} - \frac{d_M}{2} \right) + I_{M_{i-1}}^j \left(\frac{D_M}{2(\Delta x)^2} \right) + \frac{b\beta}{H} \left(\frac{S_{M_i}^j + S_{M_i}^{j+1}}{2} \right) \left(\frac{I_{H_i}^j + I_{H_i}^{j+1}}{2} \right) \quad (8)$$

Equations (5-8) form an algebraic system; hence the system is arranged in matrix form and solved simultaneously to obtain the numerical results. The initial data for susceptible and infectious compartments for both human and mosquito populations are polynomial functions. To obtain the actual density of infectious humans, actual cases in Song, Kapit, Belaga and Marudi can be obtained by dividing the prevalence of malaria to the area of each district. Then, the distances between these locations are calculated according to the latitude and longitude; thus we get $x = (\text{Song, Kapit, Belaga, Marudi}) = (0, 43, 200, 375)$.

Polynomial functions are found to better to represent the density data for each of these locations. Hence, the initial data used are:

$$S_H(\phi, x) = -1.55 \times 10^{-7} x^3 + 0.0001453 x^2 - 0.03891 x + 5.135,$$

$$I_H(\phi, x) = -1.229 \times 10^{-11} x^3 + 8.003 \times 10^{-8} x^2 - 1.454 x + 0.001,$$

$$S_M(\psi, x) = -8.067 \times 10^{-6} x^3 + 0.003249 x^2 - 0.05605 x + 7.624,$$

$$\text{and } I_M(\psi, x) = -1.697 \times 10^{-8} x^3 + 1.193 \times 10^{-5} x^2 - 0.002807 x + 0.3,$$

$$\forall \phi \in [-\tau_H, 0], \psi \in [-\tau_M, 0], x \in [0, 375].$$

Neumann boundary condition is applied where

$$\frac{\partial S_H(0, t)}{\partial x} = 0, \frac{\partial I_H(0, t)}{\partial x} = 0, \frac{\partial S_M(0, t)}{\partial x} = 0, \frac{\partial I_M(0, t)}{\partial x} = 0$$

$$\frac{\partial S_H(375, t)}{\partial x} = 0, \frac{\partial I_H(375, t)}{\partial x} = 0, \frac{\partial S_M(375, t)}{\partial x} = 0,$$

and

$$\frac{\partial I_M(375, t)}{\partial x} = 0.$$

Since we do not have the data for mosquito population in the districts in Sarawak, the initial data for susceptible and infectious mosquito are estimated to fit the actual prevalence. To obtain best fit, the objective is to attain a very low root mean square error (RMSE). RMSE is the difference between the predicted and actual density and the smaller it is, the better. The parameter value for mosquito such as death rate d_M is determined by finding the best fit numerical result compared to prevalence in the four districts. For mosquito death rate, d_M if we consider the average lifespan of mosquito only in ideal situation, the parameter value will be around 0.333-0.5 week⁻¹ [Kon and Labadin, to be published]. As it is not possible to measure directly the life span of mosquitoes in nature, the value is varied until we obtain the best fit curve. The values for mosquito death rate in the four locations differ and are lower than the predicted value for ideal situation. As it is challenging to decide on the value of the speed of the random movement of both human and mosquito populations, the diffusion rate for human population, D_H and mosquito population D_M are decided upon comparison of actual and simulation of the infectious human density.

Comparing Figure 1 which depicts the actual cases in Song, Kapit, Belaga and Marudi and the numerical result produced from the generic model in Figure 2, it is clear that the magnitude and behavior of the infectious human density is in good agreement for the different locations in time.

Next, we would like to compare the simulated infectious human density to that of the actual notified

cases in each of the four districts. The initial conditions used are the same except for:

$$I_M(\psi, x) = -6.857 \times 10^{-8} x^3 + 4.192 \times 10^{-5} x^2 - 0.007141 x + 0.4,$$

$$\forall \phi \in [-\tau_H, 0], \psi \in [-\tau_M, 0], x \in [0, 375].$$

This is because the change in the initial data for infectious mosquito population actually increases the accuracy of the model in producing simulated cases which are similar to actual cases. Hence, the initial condition plays a vital role in getting a good approximation. The simulated density is graphed on the same axis as the actual cases to compare them visually and the RMSE is calculated.

In Figure 3, the density of weekly actual malaria cases in Song and the predicted cases increases from week 0 to week 35. The disease prevails in the population and displays similar behavior for both the actual and simulated cases. RMSE for this particular set of data is 6.2334×10^{-4} .

The numerical result of the generic model's density of malaria infectious humans in Kapit resembles the actual notified cases closely as depicted in Figure 4. The RMSE is a low 4.5217×10^{-4} . Malaria cases surge from week 0 to week 35 and are alike in both cases. As shown in Figure 5, the actual density of infectious humans is slightly higher but both display similar growth of the disease until week 35. The RMSE is calculated to be 4.1168×10^{-4} . Finally, the actual prevalence of malaria in Marudi can be seen in Figure 5. The numerical result from the generic model shows a steady, almost linear increment while the notified cases exhibit a sharper increase up to week 5, then it continues to grow but at a slower rate. Both results agree that the disease prevails in Marudi by week 35. The RMSE for this comparison is 8.0379×10^{-4} .

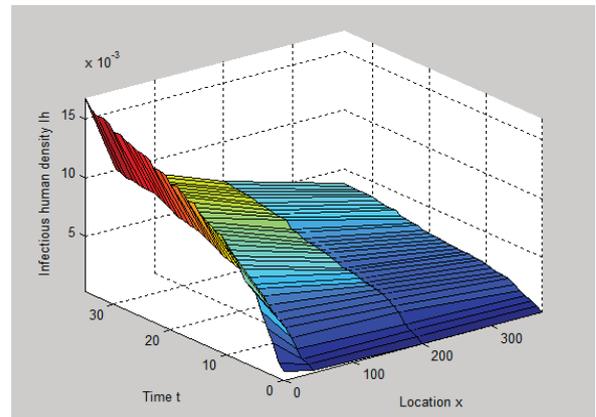


Figure 1: The actual density of infectious humans in time and space

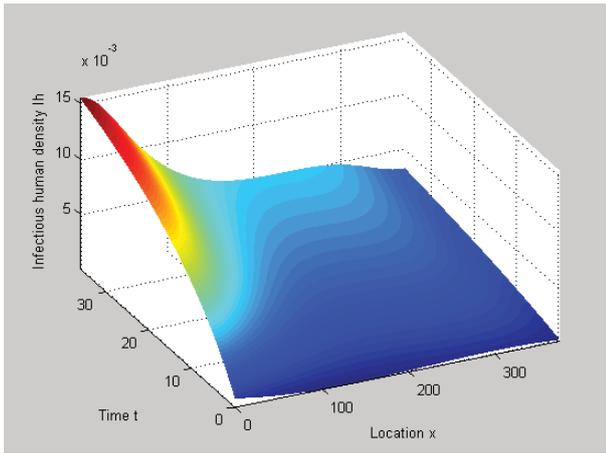


Figure 2: The simulated density of infectious humans in time and space

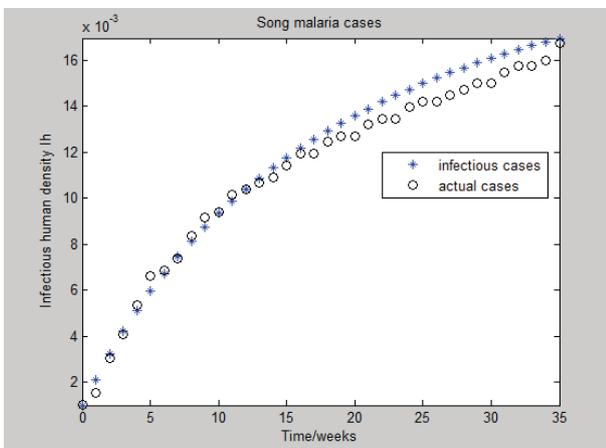


Figure 3: Weekly actual malaria cases in Song from week 39 in 2012 to week 22 in 2013 and simulated cases.

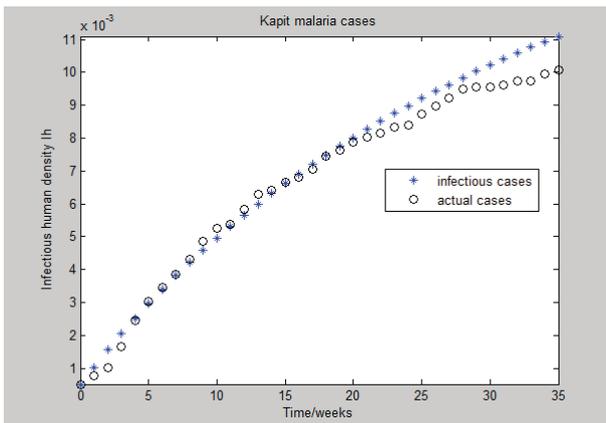


Figure 4: Weekly actual malaria cases in Kapit from week 39 in 2012 to week 22 in 2013 and simulated cases.

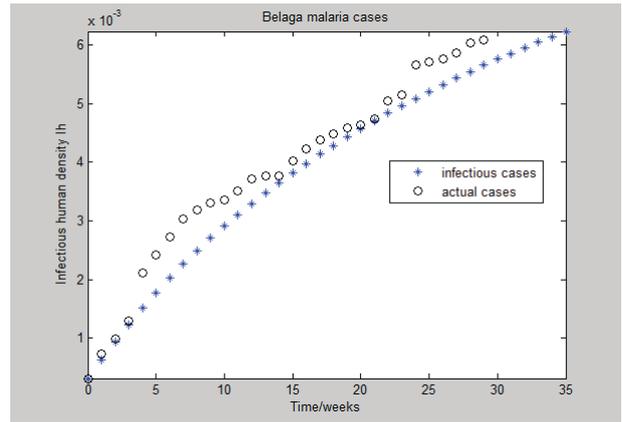


Figure 5: Weekly actual malaria cases in Belaga from week 39 in 2012 to week 22 in 2013 and simulated cases.

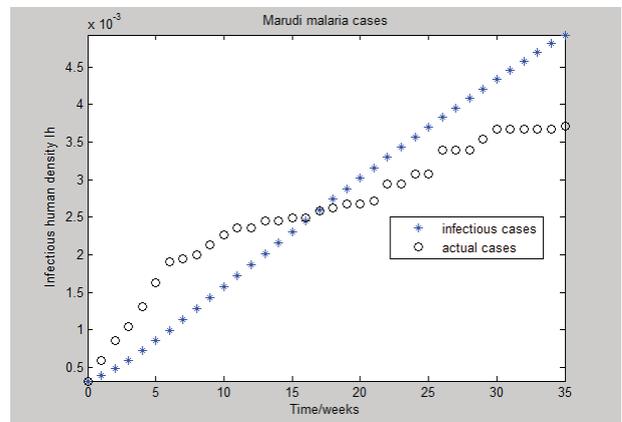


Figure 6: Weekly actual malaria cases in Marudi from week 39 in 2012 to week 22 in 2013 and simulated cases.

CONCLUSION AND FUTURE WORK

A generic model for multiple mosquito-borne diseases is discussed in this paper. This generic model is formulated to be utilized for many different types of mosquito-borne diseases. Here, we would like to use this model to produce results for malaria. Simulations from this generic model are compared with actual malaria cases in four districts in Sarawak, namely Kapit, Song, Belaga and Marudi. Thirty-six weekly notified malaria cases are obtained from Sarawak State Health Department and compared with simulations. The spatio-temporal model is found to be able to reproduce actual cases in the four locations. The disease is endemic in all four districts. The simulated cases for Song and Kapit are found to be in good agreement with that of the actual notified cases. Numerical results for Belaga and Marudi shows similar behavior, that is the disease prevails in the population although the magnitudes slightly differ. In this study, it is found that the generic model for mosquito-borne diseases is able to reproduce malaria cases which correspond to the four districts in Sarawak. For future work, we would like to study whether this generic model is able to reproduce results

for other mosquito-borne diseases such as dengue. Factors concerning spatial heterogeneity which contributes to the spread of diseases will also be identified and the optimal control measure can be determined.

ACKNOWLEDGMENTS

This work was supported by Universiti Malaysia Sarawak and Ministry of Education through the Fundamental Research Grant Scheme number FRGS/2/2013/ICT07/UNIMAS/02/6 to J.Labadin.

REFERENCES

- Anderson, R.M. and R.M. May. 1979. Population Biology of Infectious Diseases I. In *Nature* 280:361-367.
- Anderson, R.M., R.M. May, and B. Anderson. 1992. *Infectious diseases of humans: dynamics and control (Vol. 2)*. Oxford: Oxford university press.
- Aniṭa, S. and V.Capasso. 2012. "Stabilization of a reaction–diffusion system modelling a class of spatially structured epidemic systems via feedback control". In *Nonlinear Analysis: Real World Applications*, 13(2), 725-735.
- Cantrell, R.S. and C., Cosner. 2004. *Spatial ecology via reaction–diffusion equations*. John Wiley & Sons.
- Chaput, E. K.; J.I. Meek; and R. Heimer. (2002). "Spatial Analysis of human granulocytic ehrlichiosis near Lyme, Connecticut". In *Emerging Infectious Diseases*, 8(9), 943-948.
- Chitnis, N. 2005. Ph.D. thesis, Program in Applied Mathematics, University of Arizona, Tucson, AZ.
- Cliff, A.D.; P. Haggett; and D.F. Stroup. 1992. "The Geographic Structure of Measles Epidemics in the northeastern United States". In *American Journal of Epidemiology*, 136(5), 592-602.
- Cox, F.E. 2002. History of human parasitology. In *Clinical microbiology reviews*, 15(4), 595-612.
- Department of Statistics Malaysia Sarawak, *Yearbook of Statistics 2013*.
- Ding, D.; X. Wang; and X. Ding. 2012. "Global stability of multigroup dengue disease transmission model". In *Journal of Applied Mathematics*.
- Gibson, G.J. 1997. "Markov Chain Monte Carlo Methods for Fitting Spatiotemporal Stochastic Models in Plant Epidemiology". In *Appl. Statistics*, 46, 215-233.
- Kitron, U. 2000. "Risk Maps:: Transmission and Burden of Vector-borne Diseases". In *Parasitology today*, 16(8), 324-325.
- Kon, C. and J. Labadin. 2013. Reaction-diffusion generic model for mosquito-borne diseases. In *Information Technology in Asia (CITA), 2013 8th International Conference IEEE*, 1-4.
- Kon, C.M.L. and J. Labadin. 2015. "Impact of human diffusion and spatial heterogeneity on transmission dynamics of mosquito-borne diseases". In *IT in Asia (CITA), 2015 9th International Conference. IEEE*. 1-6.
- Kon, C.M.L. and J. Labadin. 2016. "Simulating the spread of malaria using a generic transmission model for mosquito-borne infectious diseases". In *2nd International Conference on Mathematical Sciences and Statistics (ICMSS2016)*. To be published.

- Labadin, J.; C.M.L. Kon; and S.F.S. Juan. 2009. "Deterministic Malaria Transmission Model with Acquired Immunity." In *Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II WCECS 2009*. San Francisco, 779-784.
- Maidana, N.A. and H. M. Yang, 2007. "A spatial model to describe the dengue propagation". In *Trends in Applied and Computational Mathematics*, 8(1), 83-92.
- Mehlhorn, H. 2001. *Encyclopedic reference of parasitology: Diseases, treatment, therapy (Vol. 2)*. Springer Science & Business Media.
- Nedelman, J. 1984. Inoculation and recovery rates in the malaria model of Dietz, Molineaux, and Thomas. In *Mathematical Biosciences*, 69(2), 209-233.
- Ross, R. The prevention of malaria (John Murray, London, 1910).
- Sarawak Health Department. 2012. *Sarawak Weekly Epid News*.
- Thomas, J. W. (1995). *Numerical Partial Differential Equations: Finite Difference Methods, Texts in Applied Mathematics*. Berlin, New York: Springer-Verlag.
- Tumwiine, J.; J.Y.T. Mugisha; and L.S. Luboobi. 2007. "A mathematical model for the dynamics of malaria in a human host and mosquito vector with temporary immunity." In *Applied Mathematics and Computation*, 189(2), 1953-1965.
- Wang, W. and X-Q. Zhao, X.Q. 2011. A nonlocal and time-delayed reaction-diffusion model of dengue transmission. In *SIAM Journal on Applied Mathematics*, 71(1), 147-168.
- Williams, H.A. and P.B. Bloland. 2002. *Malaria control during mass population movements and natural disasters*. National Academies Press.
- World Health Organization (WHO). 2015. Achieving the malaria MDG target: reversing the incidence of malaria 2000–2015.

AUTHORS BIOGRAPHIES



CYNTHIA M.L. KON is currently pursuing her PhD in Computational Science in Universiti Malaysia Sarawak. She is interested in mathematical modeling, population dynamics and modeling the spread of infectious diseases. Her present work is on spatio-temporal modeling of transmission of mosquito-borne diseases such as dengue and malaria.



JANE LABADIN is currently an Associate Professor at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS). She received her Ph.D. in Computational Mathematics specializing in Fluid Dynamics from the Imperial College of Science, Technology and Medicine, London, UK in 2002. Her Bachelor degree in Applied Mathematics was from the same university in 1995. She obtained her Master in Computation in 1997 from the University of Manchester, Institute of Science and Technology, UK. Her research interest is in computational modeling of dynamical systems.

SOME GPSS OPPORTUNITIES FOR MODELING OF TIMESTAMP ORDERING IN DDBMS AND SIMULATION INVESTIGATIONS

Svetlana Vasileva
School of Computer Science
Varna University of Management
13 A Oborishte Str, 9000, Varna, Bulgaria
E-mail: svetlanaeli@abv.bg

KEYWORDS

Simulation models, distributed transactions, distributed timestamp ordering, distributed 2PL, voting.

ABSTRACT

This paper considers some opportunities of the system for simulation modelling GPSS World Personal Version for simulating algorithms for transaction concurrency control (CC) in Distributed database management systems (DDBMS). Models of Timestamp ordering (TSO) algorithm and Two-version Two-phase locking (2V2PL) in DDBMS are presented. Both approaches – two version data and timestamps (and others) are used in database management systems for avoiding transaction deadlock. Method of TSO and method of 2V2PL are still not investigated enough. However, the use of timestamps makes the CC algorithms more complex due to the restarting of the transactions from service and the additional waiting for processing. Therefore results of the implementation the simulating algorithms are showed in comparative view.

INTRODUCTION

Simulation modeling is becoming more widespread and used as system-and an extremely valuable link in the process of decision-making, so use with other software systems for making a decision in the systems to support decision making. Nowadays the systems for modeling is a powerful analytical tool in which they integrated the all newest information technologies, for the purpose of constructing models and interpretations of simulating results, multimedia and video, supporting animation in real time, object-oriented programing, Internet-solutions, etc.

The paper considers one of the most famous and universal systems for simulation modeling – GPSS World, and especially its opportunities for simulating concurrency control algorithms in distributed database systems and making simulation investigations on the implementation of these algorithms. Simulation modeling allows us to explore queuing systems to different types of input flows and intensities of arrival of requests at the entrances of systems and determine the main features of the same. All these and other specific characteristics of GPSS World make it possible to develop simulation models of a variety of simple and

complex systems running on different algorithms and to carry out various types of research. Such studies are an analysis of concurrency control (CC) of distributed transactions in distributed database management systems as represented in (Culciar and Vasileva 2015). In the cited work are presented some simulation studies of the implementation of Centralized Two-Phase Locking (2PL) in Distributed database management systems (DDBMS). The paper presents a model of another approach for transaction CC in DDBMS named Timestamp ordering (Connoly and Begg 2002) and some simulation results also.

ADVANTAGES OF SIMULATION MODELING

Among the advantages of simulation (Tomashevskii and Zhdanova 2003) and others, we could cite the following: present only essential for understanding the behavior parts; the model can be built before the real system, with a much less resources; various parameters may change during the modeling; the model takes into account the random nature of the processes in the real system; to conduct modeling are not required deep knowledge of computational mathematics. This allows applying simulation as a universal approach to decision-making under conditions of uncertainty in the models with thinking of factors that are difficult to be formalized, but also to use the basic principles of the systematic approach to solve practical problems.

The problems that arise in simulation modeling (Tomashevskii and Zhdanova 2003; etc.), are: modeling results are always approximated; optimization of the modeled system is possible, but difficult and requires great computing power; need for validation and verification of the model and actual construction is complicated; needed is a specialist in modeling; there is always a risk to build an inadequate model. One of the purposes of the work is to try to show some GPSS World opportunities that could help us to make models, which we can trust when making a decision. These opportunities are shown on the example of the GPSS model of distributed transaction Timestamp ordering in distributed database management systems.

TIMESTAMP ORDERING AND DDBMS

One of the major problems in database management systems is concurrency control. There are mainly 3

methods for CC: protocols using 2PL, protocols using timestamping and validation check up. The first two methods have monoversion and multiversion variations. The variations of CC protocols are not only these. The 2PL protocol in DDBMS can be on Centralized 2PL, Distributed 2PL, Primary copy 2PL and Voting 2PL. (Connolly and Begg 2002). And everyone of these protocols could have monoversion, two-version and multiversion variant as presented ones in (Chardin 2005; Date 2000; etc.). Every one of the new variant of the CC protocol is needed in its study, validation and verification. That is the reason to use the simulation modeling.

In the centralized database systems the task of timestamp (TS) protocols is the global alignment of transactions so that the older transactions (which have smaller TS) in case of conflict to receive priority (Connolly and Begg 2002; Date 2000; Tanenbaum and Steen 2007; etc.). The general approach in the DDBMS is concatenation of local timestamp with the unique identifier of the node (*<local TS>*, *<node identifier>*). (Connolly and Begg 2002; etc.) The node identifier value has a smaller weight coefficient which guarantees the order of the events in accordance with the moment of their appearance.

The serving of global transaction in the distributed timestamp ordering (DTO) modeling algorithm is performed according to the algorithm of timestamps, shown in fig. 1. The schema in fig. 1 demonstrates TO algorithm in a summary, described in (Connolly and Begg 2002; etc.). The algorithm uses Tomas rules (Thomas 1979) according to which:

- The duration of service transactions has the exponential distribution with parameter m ;
- To each transaction T is assigned timestamp, denoting the time of its coming into the system and the number of the site-generator. When a transaction read/write data element, it records its TS in it.
- If a transaction T wants to update data element x :
If $TS(T) < readTS(x)$, then restart(T);
If $TS(T) < writeTS(x)$, then ignore(T);
If $TS(T) > writeTS(x)$, then execute(T).
- If a transaction wants to read data element x :
If $TS(T) < writeTS(x)$, then restart (T);
If $TS(T) > writeTS(x)$, then execute(T).

GPSS WORLD AND DISTRIBUTED TSO MODELING

GPSS blocks used in the model of Distributed TSO

In the considered GPSS model of timestamping of distributed transactions we use the following GPSS blocks (Minuteman 2010):

- Blocks for generating and cancelling of transactions: GENERATE – a main block of transactions inputting in the model; TERMINATE – to output transactions from the model.

- Blocks for management of the modeling time: ADVANCE – a block for program detention of the transactions.
- Blocks connected with devices: SEIZE – for modeling of taking an implement (device) from the transaction. In case the implement is taken, a queue is made in front of it; RELEASE - for modeling of transaction implement release by the one that has taken it.
- Blocks for multi-channel devices: ENTER – for modeling of taking one or several channels by transactions entering the block; LEAVE – for release of certain number of channels.
- Blocks for queues: QUEUE – for registration of a transaction entering a queue; DEPART – reduces the length of the particular queue with a definite number of units.
- Tables: TABULATE – the value of the argument in the table is inputted into it each time when a transaction comes into the block.
- Blocks changing the route of transaction movement: TRANSFER – a basic means of route change of a transaction; TEST – points the number of the next block for transferring the transactions in meeting/not meeting some conditions; GATE – allows to change the direction of the movement of the transactions depending on 12 logical attributes.
- Blocks changing the parameters of the transactions: ASSIGN – for appropriating number values of the transaction parameters.
- Blocks for transaction's families: SPLIT – for creating a certain number of copies of the coming transaction; ASSEMBLE – unites a given number of transactions in one family; GATHER – analogous of an ASSEMBLE block, but does not take transactions out of the model and throughputs them to the next block.
- User chains (lists): LINK – for taking a transaction out of the chain of current events and locating it in a user chain, where it waits for another transaction to take it out; UNLINK – for taking a transaction out of a user chain.

Parameters of the generated transactions in the model of Distributed TSO

The parameters of generated by GPSS transactions in the Distributed TSO model (Vasileva 2012) are the same as in (Culciar and Vasileva 2015; Vasileva and Noskov 2009). The parameters of every GPSS transaction modeling distributed transaction in DBMS, that receive their value just after they enter into the model by the block GENERATE are following (fig. 1):

- P1 – Number of the transaction. The value is a sum of System Numeric Attribute MP2 and the number of the site;
- P2 – number of the generating transaction site;
- Pnel - number of elements processed by the transaction
- Pel1 / Pel2– number of the first / second processed data element by the transaction (E11) / (E12);
- Pbl1 / Pbl2 – type of the operation over the element E11 / E12: 1 (r) – if read (E11) / (E12); 2 (w) if write (E11) / (E12);

P5 – phase of the transaction processing: it takes the value of 0 in the transaction coming in the model and after the end of the operation read/write it takes the value of 1. In the Distributed TSO model P5=2, if Ignore(T); P5=3, if Rollback(T);

P6 / P7 – number of the site where the first / second copy of the first data element E11 is stored;

P8 / P9 – number of the site-executor where the first / second copy of the second data element E12 is stored;

P11 – number of the user chain where the corresponding sub-transaction waits for the release of the copy data element.

P\$Vr – parameter that is used in making the decision about commit/rollback of transaction in Distributed TSO model: P\$Vr=0, if the transaction has not requested the element yet; P\$Vr=1, if T continues execution; P\$Vr=2, if Rollback(T).

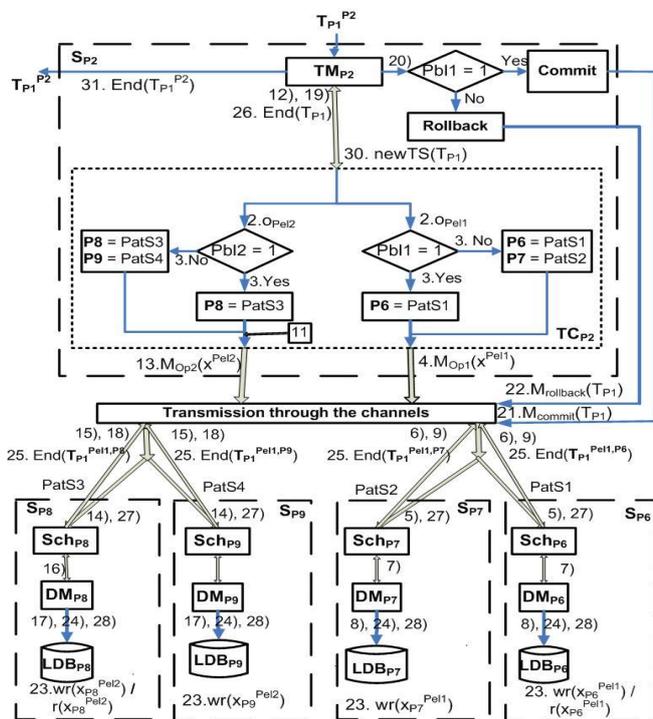


Figure 1: A scheme of distributed transaction execution by the modeling algorithm of Distributed TSO

Use of Variables in the Model of Distributed TSO

The arithmetic variables allow to calculate the arithmetic expressions which consist of operations upon SNA of the objects. Boolean variables give the opportunity to the concurrently checking of several conditions, proceeding from the assumption concerning the object's condition or the SNA values. (Minuteman 2010).

Some of the variables in the Distributed TSO model (Vasileva 2010) are:

V\$ElemN1/V\$ElemN2 – it calculates the number of the first/second element, which is processed from the transaction;

SiteRepl1/SiteRepl2 – it calculates the value of P6 and P7 / P8 and P9 (fig. 1);

RAZRBL1/RAZRBL01 – it determines whether the sub-transaction can block the first / second replica of E11 at the corresponding site-executor (P6/P7);

RAZRBL2/RAZRBL02 - it determines whether the sub-transaction can block the first / second replica of E12 in the corresponding site-executor (P8/P9);

SPIS2 / SPIS22 – it determines the number of the user chain where the sub-transaction processing the second replica of E11 / E12 can await the release of the corresponding replica.

Use of cells and matrices in the model of Distributed TSO

Cells and matrices are used for storing user numeric information. The record in these objects is used and read by the transactions. (Minuteman 2010)

In the considered model the cells are used mainly as counters:

X\$BROITR1/X\$BROITR2 - counter of transactions with length 1/2 elements;

X\$BROITR - counter of generated transactions;

X\$ZAVTR - counter of committed transactions

X\$VOT1, X\$VOT2, X\$VOT, X\$VOT12 – the counters serving in the taking of decision for continuation, ignoring or rollback of T

X\$RESTRT - counter of restarted transactions.

In the example GPSS matrices are used: for the modeling distance between the nodes in DDBMS like (Culciar and Vasileva 2015) and for modeling the distributed database (DDB) (Vasileva 2012).

MX\$RAZST and MX\$RAZDEV are used to set the mean and standard deviation of the retention time of his transactions in the transmission of messages between the nodes of the distributed database system modeled for communication costs;

GBDA1 / GBDA2 - it models the local database (LDB) for first/second copies of E11 and E12. Each row of the GBDA1 / GBDA2 corresponds to the data element from DDB. The matrices have the following columns:

- Value of the element. This value is increased by 1 when a transaction records the data element and the value decreased by 1 in cases where the transaction has written the item but then had to rollback;
- Type of operation that the current transaction carried over elements: read, write or update. When a transaction rollbacks, it writes value 0;
- The timestamp of the transaction, that is the last recorded value of the data element. In this column is written the value the P1 parameter transaction, which recorded the new value in the first column;
- The timestamp (parameter P1) of the transaction, that is the last read the value of the data element;
- The number of the site-initiated the GPSS transaction (parameter P2), that is the last processed the data element. The value is 0 if the last transaction restarted.

Use of functions in the model of Distributed TSO

The function could be a means of giving uninterrupted or discrete functional dependence between the argument of the function and its value. Functions in GPSS are assigned to a table with operators for function description. (Minuteman 2010)

In the example (Vasileva 2012) we set following functions:

XPDIS – A standard exponential distribution function that determines the exponential distribution of inflow transactions with intensity $\lambda = 1$. In blocks GENERATE we redefine the intensity of the respective transaction inflow by variants of the operand A of the statement;

DistrS1 FUNCTION V\$SiteRepl1,D6 – calculates the number of the site where is the first copy of E11 (serves to determine the value of the parameter P6);

DistrS2 FUNCTION V\$SiteRepl1,D6 - calculates the number of the site where is the second replica of E11 (serves to determine the value of the parameter P7);

DistrS3 FUNCTION V\$SiteRepl2,D6 - calculates the number of the site where is the first copy of E12 (it determines the value of the parameter P8);

DistrS4 FUNCTION V\$SiteRepl2,D6 - calculates the number of the site where is the second replica of E12 (serves to determine the value of the parameter P9);

TransCor FUNCTION P2,D6 – determines the name of GPSS Facility modeling the transaction coordinator of the current transaction;

TraMan FUNCTION P2,D6 - determines the name of the GPSS Storage Entity modeling the transaction manager of the current transaction;

Opash1 FUNCTION P6,D6 – determines the name of queue in front of facility entity modeling scheduler of first sub-transaction that processes the first copy of E11;

Opash2 FUNCTION P7,D6 - determines the name of queue in front of the scheduler of second sub-transaction that processes the second copy of E11;

Opash3 FUNCTION P8,D6 - determines the name of queue in front of facility entity modeling scheduler of third sub-transaction that processes the first copy of E12;

Opash4 FUNCTION P9,D6 - determines the name of queue in front of facility entity modeling scheduler of fourth sub-transaction that processes the second copy of E12

BrEl FUNCTION RN4,D2 – transaction length - it is calculated in number of elements processed by transaction.

Use of queues in the model of Distributed TSO

The movement of the transaction flow could be detained due to inaccessibility of the resources. In this case the transactions make a queue. There could be defined points in the model where to gather statistics about the queues (queue registrators). Then the interpreter will gather the statistics about the queues (length, average time of the stay in the queue, etc.) automatically. (Minuteman 2010; Tomashevskii and Zhdanova 2003) In this reason we set block QUEUE before every Facility entity and Storage entity (and block DEPART

after the transaction serving by corresponding Facility/Storage) included in the model. (Vasileva 2012) In order to collect statistics during the service transaction model is set statement QUEUE Totaltim after the segment setting the transaction length (parameter Pnel), the numbers of data elements (transaction parameters Pel1 (and Pel2)) processed by the transaction and the operations types (P3 (and P4)). This segment is after the generating transaction segment (Set the values of the parameters P1 and P2). And we set the statement DEPART Total Time before the block TERMINATE modeling, leaving the model by the current transaction.

Use of tables in the model of Distributed TSO

The tables serve to gather statistics about casual quantities. They consist of frequency classes in which the number of concrete quantity hits is recorded (some of the GPSS System Numeric Attributes (SNA)). (Minuteman 2010; Tomashevskii and Zhdanova 2003)

In our model and studies on it, we used these GPSS tables:

DaTable – It serves of tabulating the time of residence of every transaction in the model (GPSS SNA M1). We set the block TABULATE DaTable before the TERMINATE block by which the current transaction leaves the model.

RespTime – Table of time of residence of the GPSS transactions in the queue TotalTim.

GPSS WORLD WINDOWS AND WATCHING SIMULATIONS

The windows on the GPSS World environment, provide excellent opportunities for observing the work of the modeled systems. Choosing which windows should be open on the screen to observe the simulation is done by choosing the command Simulation Window from the Window menu (Minuteman 2010):

- Blocks Window, which gives information about: labels and names of the blocks; number of entries in the corresponding block and the others. The window allows chronological tracking of transactions in blocks at model time;

- Facilities Window – a window of single channel devices - gives information about: Number / name of the device; Number of inputs; Rate of use, Average time of residence of the transaction in the device; State of readiness; Number of the last transaction occupying device; Number of interrupted transaction in the device; Number of interrupting device transactions; Number of transaction, pending special conditions; Number of transactions, pending the holding of the device; etc. In the example we use the Matrix window.

- Matrix Window – a window of the matrices (fig. 2 shows combined the “windows” of GBDA1 and GBDA2 matrices) - shows results in values of the data elements during a simulation. We can watch the concurrent change the values of the first and second replicas of the corresponding data element (The values

in the first column are the numbers of the elements. The values in the second column are the values of the first copies and the values in the fourth column are the values of the second replicas of the data elements.). Monitoring changes in the values of the elements in the second and the fourth columns we can make conclusions whether the modeling algorithm is executed correctly (if the algorithm is correct, the values in the second column should be the same as in the fourth column).

GBDA1			GBDA2		
Elem. Z	1	2	1	2	3
1	19	0	19	0	0
2	15	0	17	0	0
3	17	0	17	0	0
4	16	0	17	0	0
5	21	0	23	0	0
6	24	0	25	0	0
7	25	0	26	0	0
8	18	0	19	0	0
9	26	0	26	0	0
10	31	0	33	0	0
11	23	0	30	0	0
12	23	0	23	0	0
13	28	0	30	0	0
14	27	0	30	0	0
15	30	0	30	0	0
16	30	0	27	0	0
17	31	0	31	0	0
18	30	0	31	0	0
19	29	0	29	0	0
20	29	0	29	0	0
21	37	0	37	0	0
22	36	0	36	0	0
23	34	0	34	0	0
24	33	0	36	0	0
25	37	0	39	0	0
26	45	0	43	0	0
27	38	0	38	0	0
28	45	0	46	0	0
29	43	0	43	0	0
30	45	0	45	0	0
31	44	0	44	0	0
32	44	0	48	0	0

Figure 2: Combined view to the windows of matrices GBDA1 and GBDA2 for demonstration and tracing of the transaction service on Distributed TSO

Fig. 3 shows a combined window to monitor the parallel execution of transactions by monitoring the change of values in the lock tables of the copies of the data elements in simulation Distributed 2V2PL (Vasileva and Noskov 2009): first column – the numbers of the elements; second column – the lock types of the first replicas; third column – the numbers of transactions locked the first replicas; fourth column – the lock types of the second replicas; fifth column – the numbers of transactions locked the second replicas.

LTA1			LTA2		
Elem. Z	1	2	1	2	3
1	0	1280.663	0	1280.663	0
2	0	1220.201	0	1220.201	0
3	0	1240.278	0	1050.066	0
4	0	1150.828	0	1140.070	0
5	0	126.711	0	126.711	0
6	0	1270.203	0	1270.203	0
7	0	1190.707	0	1190.056	0
8	0	1220.536	0	1220.536	0
9	0	1240.842	0	1140.382	0
10	0	1270.447	0	1160.813	0
11	0	126.879	0	126.879	0
12	0	1280.266	0	1280.266	0
13	0	1200.259	0	1200.259	0
14	0	1200.786	0	1200.786	0
15	0	1300.309	0	1240.417	0
16	3	1270.821	0	1270.821	0
17	0	1270.203	0	1270.203	0
18	0	1280.879	0	1170.191	0
19	0	1240.566	0	1240.566	0
20	0	1240.139	0	1240.139	0
21	0	1270.447	0	1200.769	0
22	0	125.953	0	125.953	0
23	0	1190.123	0	1000.570	0
24	0	1270.396	0	1270.396	0
25	0	125.953	0	125.953	0
26	0	1200.614	0	1200.614	0
27	0	1190.044	0	1190.044	0
28	0	1200.223	0	1200.223	0
29	0	1250.798	0	1250.798	0
30	0	1250.112	0	1250.112	0
31	0	1250.845	0	1250.845	0
32	0	1290.700	0	1290.700	0

Figure 3: Windows of matrices LTA1 and LTA2 for tracing of the transaction service on Distributed 2V2PL

- Table Window – a window of the tables – a diagram of the frequency distribution of the tabulated transactions. (fig. 4 and fig. 5). Several windows can be open and ordered on the screen in the demonstration of a model and different aspects and elements of the modeled system can be watched in them.

We can observe the frequency distribution of the tabulated transactions as during the simulation and after

modeling - finalized and this final version of the charts can be compared with published benchmarks (or be compared with other reported graphics as we could do for fig. 4 and fig. 5).

Fig. 4 and fig. 5 show frequency distribution of Response time (RT) of transactions. Frequency distribution of RT is another indicator of the performance of concurrency control algorithms. The diagrams of Frequency distribution of RT are built automatically by the formulated in the GPSS model tables (tables named DaTable in the Distributed TSO and Distributed 2V2PL models). On fig. 4 is demonstrated the histogram of Frequency distribution of RT in modeling Distributed TSO at the total intensity of the input streams 100 tr/s (maximum load on the system) and observation time 28.8 seconds. Fig. 5 shows the results on the same conditions, but for the Distributed 2V2PL model.

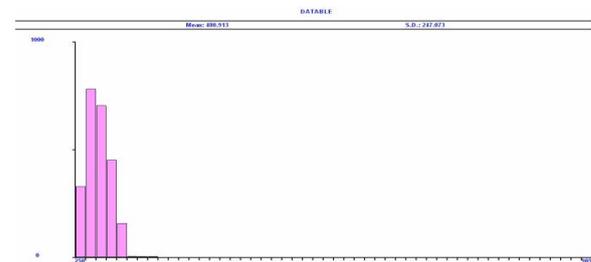


Figure 4: Frequency Distribution of RT in modeling Distributed TSO at $\lambda = 100$ tr/s

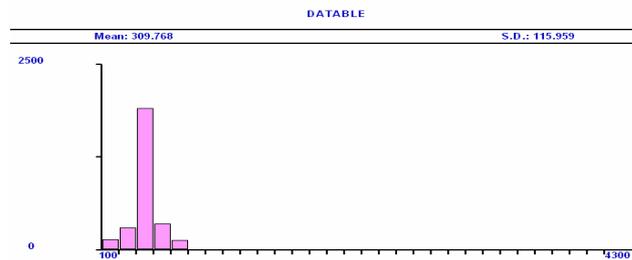


Figure 5: Frequency Distribution of RT in modeling Distributed 2V2PL at $\lambda = 100$ tr/s

The tables of Frequency distribution of RT besides that serve comparative analysis of concurrency control algorithms, serve also to assess the reliability of modeling algorithms by comparing with the template chart of Frequency distribution of RT (TPC Council 2010).

Similarly, can be compared with the template graphics and charts for throughput of fig. 6, fig. 7 and fig. 8.

SIMULATION RESULTS

Through defined in the model tables and reports generated after modeling, we can determine one of the most important features for DDBMS - the Response time. With defined in the model variables and cells, as described in Section “GPSS World and Distributed TSO Modeling” we could calculate another main

characteristics of service transactions in DDBMS: Throughput (TP) and Service Probability (SP).

Throughput of one system is calculated in the number of requests serviced per unit time (Tomashevskii and Zhdanova 2003). For our model they are respectively the values of the cell X\$ZAVTR and time modeling at different startups of the modeling algorithm.

Time modeling in the modeling algorithms is set in milliseconds in block GENERATE at the end of the models. All streams transactions are received upon an exponential law with a variable at different studies with an average length of the interval. In all modeling algorithms we consider 6 streams generated by GPSS transactions modeling 6 sites in distributed database system, from which Poisson law shall go global transactions.

The diagram of fig. 6 presents the results of simulations of Throughput for Distributed 2V2PL and Distributed TSO algorithms at the same intensities of input flows depending on the monitoring period (in seconds): The graph marked with a thin blue dashed line (2V2PL) and the graph indicated by a thick black line and square markers (TSO) – 6 streams, each with an average intensity 4,17 tr/s (minimum load - intensity cumulative flow 25 tr/s and operand); The graph marked with thin black line and asterisks (2V2PL) and in the graph illustrated by dashed lines with triangular markers (TSO) - 6 streams, each with an intensity of 8,33 tr/s (average load - intensity of the aggregate stream 50 tr/s); The graph indicated by the thin dotted line (2V2PL) and the graph indicated by a thin blue line (2PL TSwd) - 6 streams of medium intensity 16,67 tr/s (maximum load - intensity of aggregate stream 100 tr/s).

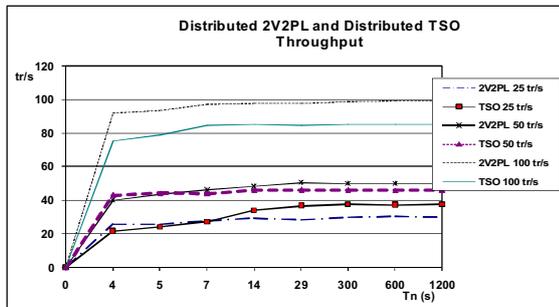


Figure 6: Throughput of the systems

To get the appropriate intensity of each of inflows in operand A in the corresponding block GENERATE we set value: 60 - maximum load; 120 - average load; 240 - minimum load. And to change the distance between the nodes of the DDBMS and conduct research on the dependence of the throughput of the system (and transaction SP and other performance indicators of concurrency control algorithms) of the distances between nodes in different modes and CC algorithms, we should change the values in the cells of the MX\$RAZST matrix (and the cells of the MX\$RAZDEV matrix).

On fig. 7 graphics are given the values that are obtained for TP by substituting the fixed in the receiving reports values of the cell X\$ZAVTR. Intensity of inflows transactions are the same as the graphs of fig. 6, it was changed only the distance matrix MX\$RAZST – their cell values are increased twice compared to models whose results are reported in the graphs of fig. 6.

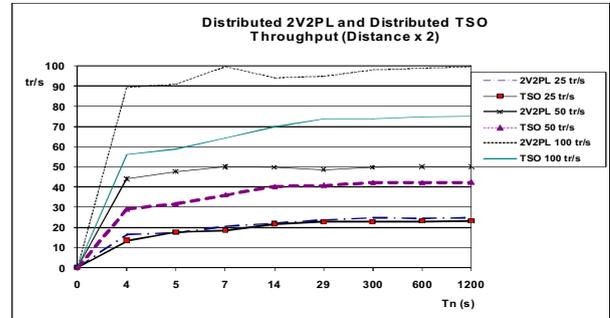


Figure 7: Throughput in the models at doubled distance between the sites in the system

From the graphs of fig. 7 and fig. 8 it can be concluded that with the increase of the distance between sites in the system, the throughput graphs are "spaced apart" more. This is very evident in the graphs at maximum load of the system (intensity of inflow 100 tr/s).

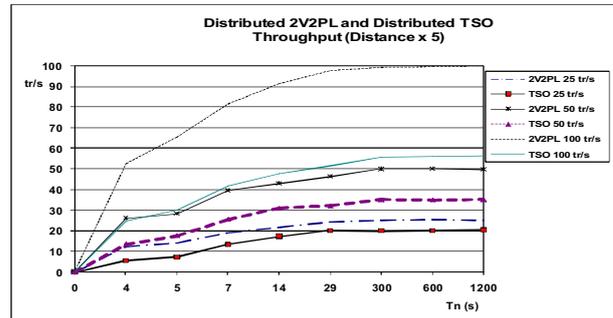


Figure 8: Throughput of the system (distance x 5)

Service probability factor or completion of service transactions serves to assess the dynamic properties of DDBMS. The probability of service P_s of distributed transactions is calculated by the formula (1) (Tomashevskii and Zhdanova 2003):

$$P_s = \frac{N_c}{N_g} \quad (1)$$

where N_c is the total number of fixed transactions (cell value of X\$ZAVTR after modeling, and N_g is the total number of transactions generated for the same period of time (cell value of X\$BROITR after modeling).

Fig. 9 presents the results for the service probability of distributed transactions at simulation algorithm Distributed TSO and Distributed 2V2PL at the same intensities of inflows (as for Figure 6).

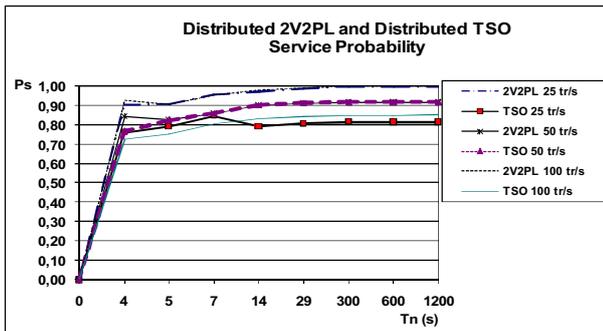


Figure 9: Service probabilities in the 2V2PL model and the TSO model

In the diagram of Figure 10 are shown in graphical form the data collected from the reports generated after the simulations conducted and reported in (Vasileva 2012). It can be seen that the graphics of RT measurements have the same kind - a rapid increase in the beginning, slow growth and stationary mode.

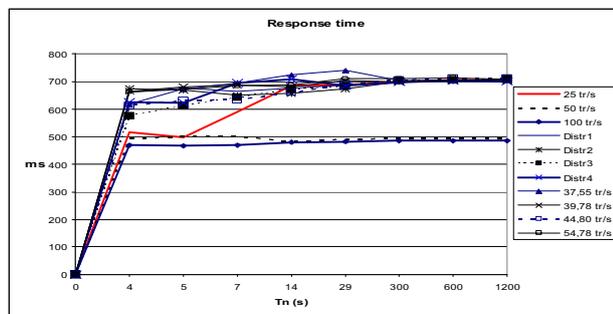


Figure 9: Response time of the Distributed TSO model in the different intensities of the incoming streams

CONCLUSION

The system of simulation GPSS World permits creation of effective simulation models of transaction concurrency control (in particular model of distributed transaction Timestamp ordering in DDBMS).

The demonstrated model developed in the simulation GPSS environment proves the great opportunities of the GPSS for simulation of algorithms for concurrency control of the distributed transactions in DDBMS. Such systems and the work algorithms of their nodes are part of the enormous field of discrete event systems.

The advantages of modeling such systems with GPSS (generating various flows of different types of requests with different in type and number parameters, realization of the service algorithm, branching out (even cyclic recurrence)) are shown in the illustrated model.

The developed simulation models (as well as their complicated analogues – for longer transactions which process two and more data elements) allow the needed statistics to be gathered with the purpose to investigate and analyze the protocols for timestamp ordering and two-phase locking in DDBMS. On the basis of the gathered data from the simulations we define the

coefficients which serve to assess the effectiveness of the algorithms for concurrency control in DDBMS.

REFERENCES

- Chardin, P. 2005. "Data multiversion and transaction concurrency control". <http://citforum.ru/database/articles/multiversion>.
- Connolly, T., C. Begg. 2002. "Database Systems". Addison – Wesley.
- Culciar, A., S. Vasileva. 2015. "Simulation studies of the implementation of Centralized two-phase locking in DDBMS". In *Proceedings of the 29th European Conference on Modelling and Simulation ECMS 2015*, May 26th – May 29th, 2015, Albena (Varna), Bulgaria, 107-114.
- Date, C. 2000. "Introduction to Database Systems". MA: Addison – Wesley.
- Minuteman Software. 2010. "GPSS World Reference Manual". <http://minutemansoftware.com/reference>.
- Tanenbaum, A., M. Steen. 2007. "Distributed systems. Principles and paradigms", Pearson Prentice Hall.
- Thomas, R. 1979. "A Majority Consensus Approach to Concurrency Control for Multiple Copy Databases". In *ACM Trans. Database Systems*, 4(2), 180-209.
- Tomashevskii, V. and E. Zhdanova. 2003. "Simulations in GPSS". Bestseller, Moscow.
- TPC Council. 2010. "TPC – C Benchmark". Standard Specification Revision 5.11. Transaction Processing Performance Council, (Feb).
- Vasileva, S. 2012. "Algorithm for Simulating Timestamp Ordering in Distributed Databases". In *Proceedings of the VI International Conference on Information Systems and GRID Technologies*, (Sofia, BG, Jun.1-3), 352-364.
- Vasileva, S., Y. Noskov. 2009. "Simulation of Distributed Two-version Two-phase Locking". In *Proceedings of the Fourth All-Russia Scientific-Practical Conference on Simulation and its Application in Science and Industry "Simulation. The Theory and Practice" (IMMOD-2009)*, Vol. II, 98-103.

AUTHOR BIOGRAPHIES



SVETLANA Zh. VASILEVA was born in Plovdiv, Bulgaria and went to the Saint-Petersburg State Electrotechnical University (LETI), where she studied Engineer Process Systems "Computer Aided Design" and obtained her degree in 1994. She has also a degree as a PhD in Informatics. She worked as a programmer (1995-1996 - EF "DeytaSoft" - Plovdiv), assistant professor of Informatics (1997-2000 - Agricultural University - Plovdiv), before moving in 2001 to the University of Shumen, College Dobrich where she was a chief assistant professor in Informatics. Currently she is a chief assistant professor at Varna University of Management. Areas of scientific interests: Control systems for distributed databases; Simulation modeling of processes and systems; Application of information technologies in education.

E-mail address is : svetlanaeli@abv.bg.

An Assessment of Pharmacological Properties of *Schinus* Essential Oils A Soft Computing Approach

José Neves
Algoritmi
Universidade do Minho
Braga, Portugal
jneves@di.uminho.pt

M. Rosário Martins
Departamento de Química
Escola de Ciências e Tecnologia
Laboratório HERCULES
Universidade de Évora, Évora, Portugal
mrm@uevora.pt

Fátima Candeias, Sílvia Arantes
Departamento de Química
Escola de Ciências e Tecnologia
Instituto de Ciências Agrárias e Ambientais Mediterrânicas
Universidade de Évora, Évora, Portugal
{mfbc, saa}@uevora.pt

Ana Piteira
Departamento de Química
Escola de Ciências e Tecnologia
Universidade de Évora, Évora, Portugal
anaisabelanaisabel14@hotmail.com

Henrique Vicente
Departamento de Química
Escola de Ciências e Tecnologia
Universidade de Évora, Évora, Portugal
Algoritmi, Universidade do Minho
hvicente@uevora.pt

KEYWORDS

Schinus spp., Essential Oils, Logic Programming, Case Base Reasoning, Knowledge Representation and Reasoning, Similarity Analysis.

ABSTRACT

Plants of genus *Schinus* are native South America and introduced in Mediterranean countries, a long time ago. Some *Schinus* species have been used in folk medicine, and *Essential Oils* of *Schinus* spp. (*EOs*) have been reported as having antimicrobial, anti-tumoural and anti-inflammatory properties. Such assets are related with the *EOs* chemical composition that depends largely on the species, the geographic and climatic region, and on the part of the plants used. Considering the difficulty to infer the pharmacological properties of *EOs* of *Schinus* species without a hard experimental setting, this work will focus on the development of an Artificial Intelligence grounded Decision Support System to predict pharmacological properties of *Schinus EO*s. The computational framework was built on top of a *Logic Programming Case Base* approach to knowledge representation and reasoning, which caters to the handling of incomplete, unknown, or even self-contradictory information. New clustering methods centered on an analysis of attribute's similarities were used to distinguish and aggregate historical data according to the context under which it was added to the Case Base, therefore enhancing the prediction process.

INTRODUCTION

Schinus L. species are trees from Anacardiaceae family characterized by pungent-smell of essential oils of their leaves and fruits. Plants of genus *Schinus* are native to South America, including approximately 29 species, and some of them have been introduced to southern Europe, including Portugal, as an ornamental plant (Bendaoud et al. 2010).

Some *Schinus* species, namely *S. molle* L., *S. terebinthifolius* Raddi and *S. longifolius* (Lindl.) Speg. are used in folk medicine to treat pathologies like rheumatism, high blood pressure, respiratory and urinary infections, or as digestive, diuretic and purgative (Duke, 2002; Atti dos Santos et al. 2010; Murray et al. 2012).

The chemical characterization of *EOs* of leaves and berries of *Schinus* spp. have been reported with the presence of different monoterpenes, sesquiterpenes and triterpenes, as secondary metabolites. However, the chemical composition of *EOs* is different according to the geographic and seasonal factors and the part of the plant used for extraction, fruit or leaves (Díaz et al. 2008; El-Massry et al. 2009; Gomes et al. 2013; Martins et al. 2014).

Some studies highlighted several biological properties of *EOs*, namely antimicrobial (El-Massry et al. 2009; Deveci et al. 2010; Martins et al. 2014), antioxidant (Díaz et al. 2008; Bendaoud et al. 2010; Martins et al. 2014), anti-tumoural (Díaz et al. 2008; Bendaoud et al.

2010), analgesic and anti-inflammatory activities (Simionatto et al. 2011; Bigliani et al. 2012), and correlated them with their biochemical structure.

Taking into account the geographical and seasonal variability of chemical composition of *Schinus EO*s and the difficulty to infer their pharmacological properties without experimental assays for each *EO*, this paper describes an intelligent support system to predict pharmacological properties of *Schinus* essential oils using a *Case Based Reasoning (CBR)* approach to problem solving (Aamodt and Plaza 1994; Richter and Weber 2013). To set the structure of the information and the associate inference mechanisms, a computational framework centered on a *Logic Programming (LP)* based approach to knowledge representation and reasoning was used. It caters to the handling of unknown, incomplete, forbidden, or even self-contradictory data, information or knowledge.

KNOWLEDGE REPRESENTATION AND REASONING

At decision times the available information is not always exact in the sense that it can be estimated values, probabilistic measures, or degrees of uncertainty. Furthermore, knowledge and belief are generally incomplete, self-contradictory, or even error sensitive, being desirable to use formal tools to deal with the problems that arise from the use of those types of data, information, or knowledge (Neves 1984; Neves et al. 2007). *Logic Programming (LP)* has been used for knowledge representation and reasoning in different areas, like Model Theory (Kakas et al. 1998; Pereira and Anh 2009), and Proof Theory (Neves 1984; Neves et al. 2007). In the present work the proof theoretical approach is followed in terms of an extension to *LP*. An *Extended Logic Program* is a finite set of clauses in the form:

$$\{$$

$$p \leftarrow p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m$$

$$? (p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m) \quad (n, m \geq 0)$$

$$\text{exception}_{p_1}$$

$$\dots$$

$$\text{exception}_{p_j} \quad (0 \leq j \leq k), k \text{ is an integer number}$$

$$\} :: \text{scoring}_{value}$$

where “?” is a domain atom denoting falsity, the p_i , q_j , and p are classical ground literals, i.e., either positive atoms or atoms preceded by the classical negation sign \neg (Neves 1984). Under this formalism, every program is associated with a set of abducibles (Kakas et al. 1998; Pereira and Anh 2009) given here in the form of exceptions to the extensions of the predicates that make the program. The term *scoring_{value}* stands for the relative weight of the extension of a specific *predicate* with respect to the extensions of the peers ones that make the overall program.

In order to evaluate the knowledge that can be associated to a logic program, an assessment of the *Quality-of-Information (QoI)*, given by a truth-value in the interval $[0, 1]$, that stems from the extensions of the predicates that make a program, inclusive in dynamic environments (Lucas 2003; Machado et al. 2008), was set. Indeed, the objective is to build a quantification process of *QoI* and *DoC (Degree of Confidence)*, the latter being a measure of one’s confidence that the argument values or attributes of the terms that make the extension of a given predicate, with relation to their domains, fit into a given interval (Fernandes et al. 2015). The *DoC* is evaluated as depicted in Figure 1 and computed using $DoC = \sqrt{1 - \Delta l^2}$, where Δl stands for the argument interval (set to the interval $[0, 1]$). Thus, the universe of discourse is engendered according to the information presented in the extensions of such predicates, according to productions of the type:

$$\text{predicate}_i - \bigcup_{1 \leq i \leq m} \text{clause}_j ((QoI_{x_m}, DoC_{x_m}), \dots$$

$$\dots, (QoI_{x_m}, DoC_{x_m})) :: QoI_i :: DoC_i$$

where U and m stand, respectively, for set union and the cardinality of the extension of *predicate_i*. QoI_i and DoC_i stands for themselves.

As an example, let us consider the logic program given by:

$$\{$$

$$\neg f_1 ((QoI_{x_1}, DoC_{x_1}), (QoI_{y_1}, DoC_{y_1}), (QoI_{z_1}, DoC_{z_1}))$$

$$\leftarrow \text{not} ((QoI_{x_1}, DoC_{x_1}), (QoI_{y_1}, DoC_{y_1}), (QoI_{z_1}, DoC_{z_1}))$$

$$f_1 \underbrace{((QoI_{[10, 15]}, DoC_{[10, 15]}), (QoI_{\perp}, DoC_{\perp}), (QoI_{2.5}, DoC_{2.5}))}_{\text{attribute's values}} :: QoI :: DoC$$

$$\underbrace{[0, 20] \quad [0, 9] \quad [5, 20]}_{\text{attribute's domains}}$$

$$\text{exception}_{f_1} ((QoI_3, DoC_3), (QoI_{[4, 6]}, DoC_{[4, 6]}), (QoI_{\perp}, DoC_{\perp}))$$

$$:: QoI :: DoC$$

$$\dots$$

$$\text{exception}_{f_k} ((QoI_{\perp}, DoC_{\perp}), (QoI_5, DoC_5), (QoI_{[6, 9]}, DoC_{[6, 9]}))$$

$$:: QoI :: DoC$$

$$\} :: 1 \quad (\text{once the universe of discourse is set in terms of the extension of only one predicate})$$

where \perp denotes a null value of the type unknown. It is now possible to split the abducible or exception set into the admissible clauses or terms and evaluate their QoI_i . A pictorial view of this process is given in Figure 2, as a pie chart.

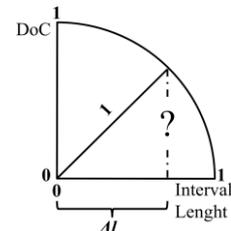


Figure 1: Evaluation of the Degree of Confidence

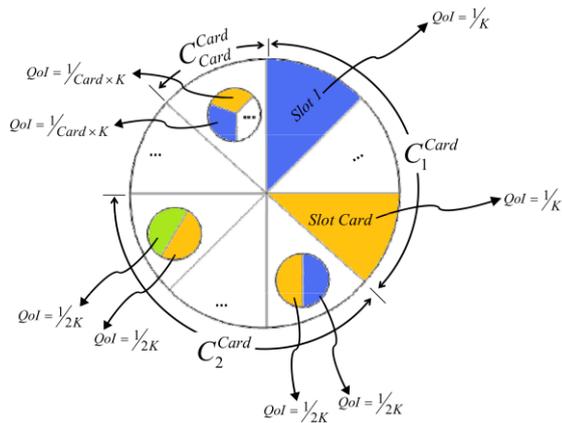


Figure 2: QoI 's values for the abducible set of clauses referred to above, where the clauses cardinality set, K , is given by the expression $C_1^{Card} + C_2^{Card} + \dots + C_{Card}^{Card}$

CASE BASED REASONING

The *CBR* methodology for problem solving stands for an act of finding and justifying the solution to a given problem based on the consideration of similar past ones, by reprocessing and/or adapting their data or knowledge (Aamodt and Plaza 1994; Richter and Weber 2013). In *CBR* – the cases – are stored in a *Case Base*, and those cases that are similar (or close) to a new one are used in the problem solving process. There are examples of its use in *The Law* with respect to *Dispute Resolution* (Carneiro et al. 2013), in *Medicine* (Janssen et al. 2014; Ying et al. 2015), among others. The typical *CBR* cycle presents the mechanism that should be followed to have a consistent model. The first stage consists of the initial description of the problem. The new case is defined and it is used to retrieve one or more cases from the repository. At this point it is important to identify the characteristics of the new problem and retrieve cases with a higher degree of similarity to it. Thereafter, a solution for the problem emerges, on the *Reuse* phase, based on the blend of the new case with the retrieved ones. The suggested solution is reused (i.e., adapted to the new case), and a solution is provided (Aamodt and Plaza 1994; Richter and Weber 2013). However, when adapting the solution it is crucial to have feedback from the user, since automatic adaptation in existing systems is almost impossible. This is the *Revise* stage, in which the suggested solution is tested by the user, allowing for its correction, adaptation and/or modification, originating the test-repaired case that sets the solution to the new problem. The test-repaired case must be correctly tested to ensure that the solution is indeed correct. Thus, one is faced with an iterative process since the solution must be tested and adapted while the result of applying that solution is inconclusive. During the *Retain* (or *Learning*) stage the case is learned and the knowledge base is updated with the new case (Aamodt and Plaza 1994; Richter and Weber 2013).

Despite promising results, the current *CBR* systems do not cover all areas, and in some cases, the user cannot choose the similarity(ies) method(s) and is required to

follow the system defined one(s), even if they do not meet their needs (Richter and Weber 2013; Neves and Vicente n.d.). But, worse than that, in real problems, access to all necessary information is not always possible, since existent *CBR* systems have limitations related to the capability of dealing, explicitly, with unknown, incomplete, and even contradictory information. To make a change, a different *CBR* cycle was induced (Figure 3). It takes into consideration the case's QoI and DoC (Neves and Vicente n.d.). It deals not only with unknown, incomplete, forbidden, and even self-contradictory data, information or knowledge, in an explicit way, but also contemplates the cases optimization in the *Case Base*, whenever they do not comply with the terms under which a given problem as to be addressed (e.g., the expected degree of confidence on the diagnostic was not attained), either using particle swarm optimization procedures (Mendes et al 2003), or genetic algorithms (Neves et al 2007), just to name a few.

METHODS

The data set was obtained based on experimental researches with EOs of leaf and fruit of *Schinus molle* collected in Alentejo (Martins et al. 2014) and *S. molle*, *S. terebinthifolius* and *S. longifolius* collected in Brazil and Argentina (Atti dos Santos et al. 2010; Gomes et al. 2013; Murray et. al 2012).

The knowledge database is specified in terms of the extensions of the relations depicted in Figure 4, which denotes a situation where one has to manage information aiming to evaluate the pharmacological properties of *Schinus* essential oils. Under this scenario some incomplete and/or unknown data is also available. For instance, in the former case, the data regarding antioxidant tests are unknown, as depicted by the symbol \perp , while the percentage of monoterpenes hydrocarbons ranges in the interval $[68, 72]$. The *Plant Part* column ranges in the interval $[0, 1]$, wherein 0 (zero), and 1 (one) denote, respectively, *leaves* and *fruit*.

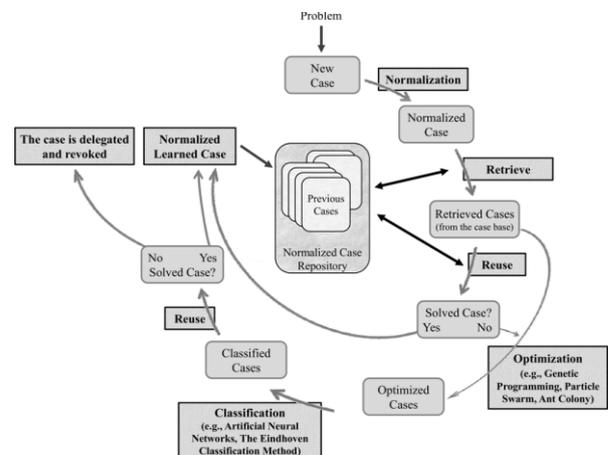


Figure 3: An extended view of the CBR cycle

		Pharmacological Activity Screening Assessment									
Attributes of the Feature Vector:	#	Plant Part	Monoterpenes Hydrocarbons	Oxygenated Monoterpenes	Sesquiterpenes Hydrocarbons	Oxygenated Sesquiterpenes	CL ₅₀	DL ₅₀	Hippocratic Screening	Antioxidant Activity	Description
Feature Vector Attributes:	1	0	[68, 72]	[0.5, 2]	[2, 4]	[13, 15]	48	2000	1	⊥	Description 1
	2	0	[63, 65]	[0.6, 0.8]	[30, 32]	[2, 3]	⊥	1500	2	2	Description 2

	73	1	[91, 98]	[0, 1]	[0, 1]	[1, 3]	67	2500	3	5	Description 73
Feature Vector Domains:		[0, 1]	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[25, 3000]	[100, 5000]	[0, 6]	[0, 12]	

Figure 4: A fragment of the knowledge base aiming to predict pharmacological activity of essential oils of *Schinus* species

Applying the algorithm presented in Fernandes (2015) to the fields that make the knowledge base for pharmacological activity screening assessment (Figure 4), excluding of such a process the *Description* ones, and looking to the DoC_s values obtained, it is possible to set the arguments of the predicate **pharmacological activity** ($pharm_{act}$) referred to below, that also denotes the objective function with respect to the problem under analysis:

$$\begin{aligned}
pharm_{act} : & P_{lant}P_{art}, M_{onoterpenes}H_{ydrocarbons}, \\
& M_{onoterpenes}O_{xigenated}, S_{esquiterpenes}H_{ydrocarbons}, \\
& S_{esquiterpenes}O_{xigenated}, CL_{50}, DL_{50}, H_{ippocratic} \\
& S_{creening}, A_{ntioxidant}A_{ctivity} \rightarrow \{0, 1\}
\end{aligned}$$

Begin %DoCs evaluation%

%The predicate's extension that sets the Universe-of-Discourse for the term under observation is fixed%

$$\begin{aligned}
& \{ \\
& \quad \neg pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad \quad \leftarrow not\ pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad pharm_{act} \left(\underbrace{((1_{\perp}, DoC_{\perp}), (1_{[85, 93]}, DoC_{[85, 93]}), \dots, (1_{\perp}, DoC_{\perp}))}_{\substack{\text{attribute's values} \\ [0, 1] \quad [0, 100] \quad \dots \quad [0, 12]}} \right) :: 1 :: DoC \\
& \quad \quad \quad \underbrace{\hspace{10em}}_{\text{attribute's domains}} \\
& \} :: 1
\end{aligned}$$

%The attribute's values ranges are rewritten%

$$\begin{aligned}
& \{ \\
& \quad \neg pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad \quad \leftarrow not\ pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad pharm_{act} \left(\underbrace{((1_{[1, 1]}, DoC_{[1, 1]}), (1_{[85, 93]}, DoC_{[85, 93]}), \dots, (1_{[0, 12]}, DoC_{[0, 12]}))}_{\substack{\text{attribute's values ranges} \\ [0, 1] \quad [0, 100] \quad \dots \quad [0, 12]}} \right) :: 1 :: DoC \\
& \quad \quad \quad \underbrace{\hspace{10em}}_{\text{attribute's domains}} \\
& \} :: 1
\end{aligned}$$

%The attribute's boundaries are set to the interval [0, 1]%

$$\begin{aligned}
& \{ \\
& \quad \neg pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad \quad \leftarrow not\ pharm_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA})) \\
& \quad pharm_{act} \left(\underbrace{((1_{[1, 1]}, DoC_{[1, 1]}), (1_{[0.85, 0.93]}, DoC_{[0.85, 0.93]}), \dots, (1_{[0, 1]}, DoC_{[0, 1]}))}_{\substack{\text{attribute's values ranges once normalized} \\ [0, 1] \quad [0, 1] \quad \dots \quad [0, 1]}} \right) :: 1 :: DoC \\
& \quad \quad \quad \underbrace{\hspace{10em}}_{\text{attribute's domains once normalized}} \\
& \} :: 1
\end{aligned}$$

where 0 (zero) and 1 (one) denote, respectively, the truth values *false* and *true*.

Exemplifying the application of the algorithm presented in Fernandes (2015), to a term (case) that presents feature vector ($P_{lant} P_{art} = 1$, $M_{onoterpenes} H_{ydrocarbons} = [85, 93]$, $M_{onoterpenes} O_{xigenated} = [1.2, 2.4]$, $S_{esquiterpenes} H_{ydrocarbons} = [0.8, 3.2]$, $S_{esquiterpenes} O_{xigenated} = [1.2, 4.3]$, $CL_{50} = 63$, $DL_{50} = 2400$, $H_{ippocratic} S_{creening} = 2$, $A_{ntioxidant} A_{ctivity} = \perp$), and applying the procedure referred to above, one may get:

%The DoC's values are evaluated%

{
 $\neg \text{pharm}_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA}))$
 $\leftarrow \text{not pharm}_{act} ((QoI_{PP}, DoC_{PP}), (QoI_{MH}, DoC_{MH}), \dots, (QoI_{AA}, DoC_{AA}))$
 $\text{pharm}_{act} \left(\underbrace{(1, 0), (1, 0.997), \dots, (1, 0)}_{\substack{\text{attribute's quality-of-information} \\ \text{and respective confidence values} \\ [1, 1] [0.85, 0.93] \dots, [0, 1] \\ \text{attribute's values ranges once normalized} \\ [0, 1] [0, 1] \dots, [0, 1] \\ \text{attribute's domains once normalized}}} \right) :: 1 :: 0.89$
 } :: 1

End.

SOFT COMPUTING APPROACH

A soft computing approach to model the universe of discourse based on CBR methodology for problem solving is now set. Indeed, contrasting with other problem solving methodologies (e.g., *Decision Trees* or *Artificial Neural Networks*), in a CBR based methodology relatively little work is done offline. Undeniably, in almost all the situations the work is performed at query time. The main difference between this new approach and the typical CBR one relies on the fact that not only all the cases have their arguments set in the interval [0, 1], but it also caters for the handling of incomplete, unknown, or even self-contradictory data or knowledge (Neves and Vicente n.d.). Thus, the classic CBR cycle was changed (Figure 3), being the *Case Base* given in terms of triples that follow the pattern:

$Case = \{ \langle Raw_{data}, Normalized_{data}, Description_{data} \rangle \}$

where Raw_{data} and $Normalized_{case}$ stand for themselves, and $Description_{data}$ is made on a set of strings or even in free text, which may be analyzed with string similarity algorithms. When confronted with a new case, the system is able to retrieve all cases that meet such a structure and optimize such a population, i.e., it considers the attributes *DoC*'s value of each case or of their optimized counterparts when analysing similarities among them. Thus, under the occurrence of a new case, the goal is to find similar cases in the *Case Base*. Having this in mind, the algorithm given in Fernandes (2015) is applied to a new case that presents feature vector ($P_{lant} P_{art} = 0$, $M_{monoterpenes} H_{hydrocarbons} = [68, 71]$, $M_{monoterpenes} O_{xigenated} = [0.8, 2.1]$, $S_{esquiterpenes} H_{hydrocarbons} = [2, 5.2]$, $S_{esquiterpenes} O_{xigenated} = [14, 16]$, $CL_{50} = 45$, $DL_{50} = 2200$, $H_{ippocratic} S_{creening} = \perp$, $A_{ntioxidant} A_{ctivity} = 3$, $Description = Description_{new}$), with the results:

$\underbrace{\text{pharm}_{act_{new}}((1, 1), (1, 0.99), \dots, (1, 1))}_{\text{new case}} :: 1 :: 0.88$

The *new case* can be depicted on the *Cartesian Plane* in terms of its *QoI* and *DoC*, and through clustering techniques, it is feasible to identify the clusters that

intermingle with the new one (symbolized as a square in Figure 5). The *new case* is compared with every retrieved case from the cluster using a similarity function *sim*, given in terms of the average of the modulus of the arithmetic difference between the arguments of each case of the selected cluster and those of the *new case* (once *Description* stands for free text, its analysis is excluded at this stage). Thus, one may have:

$\text{pharm}_{act_1}((1, 1), (1, 0.92), \dots, (1, 0)) :: 1 :: 0.85$
 $\text{pharm}_{act_2}((1, 1), (1, 0.97), \dots, (1, 0)) :: 1 :: 0.82$
 \vdots
 $\underbrace{\text{pharm}_{act_j}((1, 1), (1, 0.99), \dots, (1, 1))}_{\text{normalized cases from retrieved cluster}} :: 1 :: 0.91$

Assuming that every attribute has equal weight, the dissimilarity between $\text{pharm}_{act_{new}}^{DoC}$ and the $\text{pharm}_{act_1}^{DoC}$, i.e., $\text{dissim}_{\text{pharm}_{act_{new \rightarrow 1}}}^{DoC}$, may be computed as follows:

$$\begin{aligned} \text{dissim}_{\text{pharm}_{act_{new \rightarrow 1}}}^{DoC} &= \\ &= \frac{\|1-1\| + \|0.97-0.92\| + \dots + \|1-0\|}{9} = 0.17 \end{aligned}$$

Thus, the similarity between $\text{pharm}_{act_{new \rightarrow 1}}^{DoC}$ ($\text{sim}_{\text{pharm}_{act_{new \rightarrow 1}}}^{DoC}$) is $1 - 0.17 = 0.83$. Regarding *QoI* the procedure is similar, returning $\text{sim}_{\text{pharm}_{act_{new \rightarrow 1}}}^{QoI} = 1$.

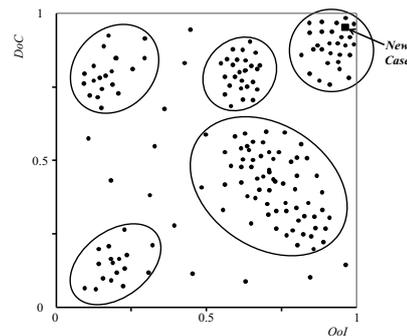


Figure 5: A case's set split into clusters

Descriptions will be compared using *String Similarity Algorithms*, in order to liken the description of the new case with the descriptions of the cases belonging to the retrieved cluster (in this study the strategy used was the *Dice Coefficient* one (Dice 1945)), with the results:

$$sim_pharm_{act_{new \rightarrow 1}}^{Description} = 0.80$$

With these similarity values it is possible to get a global similarity measure:

$$sim_pharm_{act_{new \rightarrow 1}} = \frac{0.83 + 1 + 0.80}{3} = 0.88$$

These procedures should be applied to the remaining cases of the retrieved cluster in order to obtain the most similar ones, which may stand for the possible solutions to the new problem.

A common tool to evaluate the performance of the classification models is the coincidence matrix, i.e., a matrix of size $L \times L$, where L denotes the number of possible classes (two in the present case). This matrix is created by matching the predicted and target values. Table 1 presents the coincidence matrix (the values denote the average of the 30 experiments). It shows that the model accuracy was 87.7% (64 instances of 73 correctly classified). Based on coincidence matrix it is possible to compute the sensitivity and the specificity of the model:

$$sensitivity = TP / (TP + FN) \quad (1)$$

$$specificity = TN / (TN + FP) \quad (2)$$

where TP, FN, TN and FP stand, respectively, for true positive, false negative, true negative and false positive. Briefly, sensitivity and specificity are statistical measures of the performance of a binary classifier. Sensitivity measures the proportion of true positives that are correctly identified as such, while specificity measures the proportion of true negatives that are correctly identified. In this case both metrics show values higher than 85%, (i.e., 87.0% and 88.9% for sensitivity and specificity, respectively). In addition, the *Receiver Operating Characteristic (ROC)* curves were considered. An *ROC* curve displays the trade-off between sensitivity and specificity. The *Area Under the Curve (AUC)* quantifies the overall ability of the test to discriminate between the output classes. Figure 6 depicted the *ROC* curve for the proposed model. The area under *ROC* curve is 0.88 denoting that the model exhibits a good performance in the evaluation of pharmacological properties of *Schinus* essential oils.

Table 1: The Coincidence Matrix for the ANN Model

Target	Predictive	
	True (1)	False (0)
True (1)	40	6
False (0)	3	24

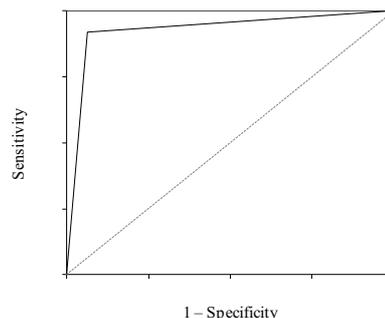


Figure 6: The *ROC* curve for the proposed model

CONCLUSIONS

This work presents an intelligent decision support system aiming to predict pharmacological activity of essential oils of *Schinus* species. It is centred on a formal framework based on *LP* for *Knowledge Representation and Reasoning*, complemented with a *CBR* approach to problem solving that caters for the handling of incomplete, unknown, or even contradictory information. Under this approach the cases' retrieval and optimization phases were heightened and the time spent on those tasks shortened in 11.3%, when compared with existing systems. The proposed approach is able to provide adequate responses since the overall accuracy was around 88% and the area under *ROC* curve is near 0.9. The proposed method allows for the analysis of free text attributes using *String Similarities Algorithms*, which fulfils a gap that is present in almost all *CBR* software tools. Additionally, under this approach the users may define the weights of the cases' attributes on-the-fly, letting them to choose the most appropriate strategy to address the problem (i.e., it gives the user the possibility to narrow the search space for similar cases at runtime).

ACKNOWLEDGMENTS

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

REFERENCES

- Aamodt, A.; and E. Plaza. 1994. "Case-based reasoning: Foundational issues, methodological variations, and system approaches." *AI Communications* 7, 39-59.
- Atti dos Santos A. C.; M. Rossato; L. A. Serafini; M. Bueno; L. B. Crippa; V. C. Sartori; E. Dellacassa; and P. Moyna. 2010. "Antifungal effect of *Schinus molle* L., Anacardiaceae, and *Schinus terebinthifolius* Raddi, Anacardiaceae, essential oils of Rio Grande do Sul." *Brazilian Journal of Pharmacognosy* 20, 154-159.
- Bendaoud, H.; M. Romdhane; J.P.Souchard; S. Cazaux; and J. Bouajila. 2010. "Chemical composition and anticancer and antioxidant activities of *Schinus molle* L. and *Schinus terebinthifolius* Raddi berries essential oils." *Journal of Food Science* 75, 466-472.

- Bigliani, M.C.; Rossetti, V.; E. Grondona; S. Lo Presti; P.M. Paglini; V. Rivero; M.P. Zunino; and A.A. Ponce. 2012. "Chemical compositions and properties of *Schinus areira* L. essential oil on airway inflammation and cardiovascular system of mice and rabbits." *Food and Chemical Toxicology* 50, 2282-2288.
- Carneiro, D.; P. Novais; F. Andrade; J. Zeleznikow; and J. Neves. 2013. "Using Case-Based Reasoning and Principled Negotiation to provide decision support for dispute resolution." *Knowledge and Information Systems* 36, 789-826.
- Deveci, O.; A. Sukan; N. Tuzun; and E.E.H. Kocabas. 2010. "Chemical composition, repellent and antimicrobial activity of *Schinus molle* L." *Journal of Medicinal Plants Research* 4, 2211-2216.
- Díaz, C.; S. Quesada; O. Brenes; G. Aguilar; and J.F. Ciccio. 2008. "Chemical composition of *Schinus molle* essential oil and its cytotoxic activity on tumor cell lines." *Natural Product Research* 22, 1521-1534.
- Dice, L. 1945. "Measures of the Amount of Ecologic Association between Species." *Ecology* 26, 297-302.
- Duke J. 2002. *Handbook of Medicinal Herbs*, CRC Press, Florida.
- El-Massry, K.F.; A.H. Ghorab; H.A. Shaaban; and T. Shibamoto. 2009. "Chemical compositions and antioxidant/ antimicrobial activities of various samples prepared from *Schinus terebinthifolius* leaves cultivated in Egypt." *Journal of Agricultural and Food Chemistry* 57, 5265-5270.
- Fernandes, F.; H. Vicente; A. Abelha; J. Machado; P. Novais; and J. Neves. 2015. "Artificial Neural Networks in Diabetes Control". In *Proceedings of the 2015 Science and Information Conference (SAI 2015)*, IEEE Edition, 362-370.
- Gomes, V.; G. Agostini; F. Agostini; A. C. Atti dos Santos; and M. Rossato. 2013. "Variation in the essential oils composition in Brazilian populations of *Schinus molle* L. (Anacardiaceae)." *Biochemical Systematics and Ecology* 48, 222-227.
- Janssen, R.; P. Spronck; and A. Arntz. 2014. "Case-based reasoning for predicting the success of therapy." *Expert Systems* 32, 165-177.
- Kakas, A.; R. Kowalski; and F. Toni. 1998. "The role of abduction in logic programming". In *Handbook of Logic in Artificial Intelligence and Logic Programming*, D. Gabbay, C. Hogger and I. Robinson (Eds.). Vol. 5, Oxford University Press, Oxford, 235-324.
- Lucas, P. 2003. "Quality checking of medical guidelines through logical abduction". In *Proceedings of AI-2003 (Research and Developments in Intelligent Systems XX)*, F. Coenen, A. Preece and A. Mackintosh (Eds.). Springer, London, 309-321.
- Machado J.; A. Abelha; P. Novais; J. Neves; and J. Neves. 2008. "Quality of service in healthcare units". In *Proceedings of the ESM 2008*, C. Bertelle, and A. Ayeshe (Eds.). Eurosis, Ghent, 291-298.
- Martins, M.R.; S. Arantes; F. Candeias; M.T. Tinoco; and J. Cruz-Morais. 2014. "Antioxidant, antimicrobial and toxicological properties of *Schinus molle* L. essential oils." *Journal of Ethnopharmacology* 151, 485-492.
- Mendes, R.; J. Kennedy; and J. Neves. 2003. "Watch thy neighbor or how the swarm can learn from its environment". In *Proceedings of the 2003 IEEE Swarm Intelligence Symposium (SIS'03)*, IEEE Edition, 88-94.
- Murray A.P.; S.A. Rodriguez; and N.P. Alza. 2012. "Chemical Constituents and Biological Activities of Plants from the Genus *Schinus*". In *Recent Progress in Medicinal Plants*, J.N. Govil (Ed.). Ethnomedicine and Therapeutic Validation, Vol. 32, Studium Press LLC, Texas, 261-287.
- Neves, J. 1984. "A logic interpreter to handle time and negation in logic databases". In *Proceedings of the 1984 annual conference of the ACM on The Fifth Generation Challenge*, R.L. Muller and J.J. Pottmyer (Eds.). ACM, New York, 50-54.
- Neves, J.; and H. Vicente. (n. d.) "A Quantum approach to Case-Based Reasoning." (In preparation).
- Neves, J.; J. Machado; C. Analide; A. Abelha; and L. Brito. 2007. "The halt condition in genetic programming". In *Progress in Artificial Intelligence*, J. Neves, M.F. Santos and J. Machado (Eds.). Lecture Notes in Artificial Intelligence, Vol. 4874, Springer, Berlin, 160-169.
- Pereira, L.M. and H.T. Anh. 2009. "Evolution prospection". In *New Advances in Intelligent Decision Technologies – Results of the First KES International Symposium IDT 2009*, K. Nakamatsu (Ed.). Studies in Computational Intelligence, Vol. 199, Springer, Berlin, 51-64.
- Richter, M.; and R. Weber. 2013. *Case-Based Reasoning: A Textbook*. Springer, Berlin.
- Simionatto, E.; M. Chagas; M. Peres; S. Hess; C. Silva; N. Ré-Poppi; S. Gebara; J. Corsino; F. Morel; C. Stuker; M. Matos; and J. Carvalho. 2011. "Chemical composition and biological activities of leaves essential oil from *Schinus molle* (Anacardiaceae)." *Journal of Essential Oil Bearing Plants* 14, 590-599.
- Ying, S.; C. Joël; J. Arnelle; and L. Kai. 2015. "Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system." *Journal of Biomedical Informatics* 56, 307-317.

TRUCK ARRIVAL MANAGEMENT AT MARITIME CONTAINER TERMINALS

Daniela Ambrosino
Lorenzo Peirano
Dept. of Economics and Business Studies (DIEC)
University of Genova
Via Vivaldi 2, 16126 Genova, Italy
E-mail: daniela.ambrosino@economia.unige.it
lorenzo.peiran@gmail.com

KEYWORDS

Truck appointment system, container terminal, mixed integer linear programming model.

ABSTRACT

In this work the management of truck arrivals in a maritime terminal is investigated as possible strategy to be used for obtaining a reduction in congestion and gate queues. In particular, a non-mandatory Truck Appointment System (TAS) is considered.

Inspired by the work of Zehedner and Feillet 2013, we propose a multi-commodity network flow model for representing a maritime terminal. We solve a mixed integer linear programming model based on the network flow for determining the number of appointments to offer for each time window in such a way to serve trucks in the shortest time as possible, thus granting trucks a “good” service level. Some preliminary results are presented. Solutions show the effectiveness of the proposed model in flattening the arrivals distribution of vehicles.

INTRODUCTION

Container maritime terminals play an important role in the logistic networks and have to be efficient intermodal nodes. In the last years mega vessel containerships have been used in order to reach economies of scale by transporting even larger numbers of containers. Thus new problems arise in the container terminals: the need of unloading and loading more containers in even smaller amounts of time requires new management strategies for avoiding congestion.

The efficient management of import and export flows is fundamental for both the terminals and the collectivity. Infact, congestion inside and outside the terminal may cause serious environmental traffic problems, besides a limitation of the efficiency of the terminals themselves. Two strategies can be used for obtaining a reduction in congestion and gate queues: the first strategy is the extension of the gate capacity, generally associated to an extension of the area of the terminal; the second strategy concerns the Truck Arrival Management (TAM).

Strategies that are not related to TAM are, for example, in Dekker et al. 2013, in which a chassis exchange system

is described, in Cullinane and Wilmsmeier 2011, in which dry ports strategies are suggested as strategic choice for maritime terminals aimed at reducing the traffic on the roads and moving it onto the rail networks. Dry ports strategies are particularly useful when terminals are close to urban and suburban areas, characterized by heavy traffic (Roso et al. 2009). In Ambrosino and Sciomachen 2014 the problem of locating dry ports for freight mobility in intermodal networks is faced.

TAM strategies aim at obtaining an efficient usage of resources inside the terminal, good service time to clients, no congestion inside and outside the terminal. Alessandri et al. 2008 propose a model for determining the best allocation of the terminal resources in such a way to optimize some key performance indexes.

TAM strategies are generally based on truck appointment systems. In fact, usually, for what concerns the number of trucks approaching the terminal, there are some picks in particular hours of a day (Guan and Liu 2009), and it is obvious that a better distribution in the arrival of trucks at the terminal, will cause smaller queues and will increase the efficiency of the terminal.

A Truck Appointment System (TAS) defines a maximum number of trucks that can approach the terminal and pass the gate during the time windows in which the working day is split.

Truck appointment systems have been introduced by some terminals in order to balance truck arrivals, such as Vancouver, Los Angeles and Long Beach (Morais and Lord 2006). The performances of these TAS are not uniform, thus suggesting that it is necessary to implement different TAS in accordance with specific local conditions (Chen et al. 2013).

Huynh and Walton 2008 stress that only if there is a correct dimension of the system, adequate to the terminal size, such that an efficient usage of the resources of the terminal is permitted, it is possible to obtain advantages from TAS. The authors consider TAS an instrument for controlling the flows and optimizing the resources usage in the terminal; they propose a method for determining the maximum number of vehicles that, in each time window, can be accepted in each zone of the terminal.

Moreover, their study takes into account the delays and the non-arrivals, searching for robust solutions. To the authors' knowledge, only another very recent work takes into account the problem of possible truck arrival deviation from the schedule in the appointment system: in Li et al. 2016 a set of response strategies for neutralizing the impact of disruptions is presented.

Among the papers investigating the problem of how to define the appointment quota for each window, in Zhang et al. 2013 an optimization model for determining these quotas while minimizing the waiting time at the gate queues and the yard waiting time, is proposed.

Chen et al. 2013 propose a time window control program to alleviate gate congestion, that is based on three steps: *i)* estimate truck arrivals based on the time window assignment and the distribution pattern of truck arrivals *ii)* estimate truck queue length using a non-stationary queuing model *iii)* optimize time window in such a way to minimize the total system cost, i.e. the truck waiting time, the idling fuel consumption, the cargo storage time and the storage yard fee.

In Zehedner and Feillet 2013 the authors evaluate the impact of the TAS in the port of Marseille that uses straddle carriers (SC) to serve trucks, trains, barges, and vessels. They present a minimum cost multi-commodity network flow model in which each commodity represents a container flow from/to a transport mode. The model simultaneously determines the number of truck appointments to accept and the number of straddle carriers to allocate to different transport modes with the objective of reducing overall delays at the terminal.

In Chen et al. 2013b the problem of sizing each window has been solved by minimizing the total number of shifted arrivals and the total truck waiting time. The authors show that good results in terms of truck idling emission can be obtained by shifting also a small number of trucks from peak to off-peak periods.

In Chen et al. 2011 another strategy used to obtain a different distribution of truck arrivals, based on time dependent tool pricing, is described. The authors propose a method that firstly determines the arrival distribution d^* that minimizes the total queues time and the disadvantages for trucks. Secondly, they try to define the pricing tools able to modify the trucks' behavior and to obtain a truck arrival process equal to d^* , while minimizing the average price paid by trucks.

In Phan and Kin 2015 the focus is on the importance of defining terminals strategies to reduce congestion by including also decisions and requirements of trucking companies. A mathematical model to make the appointment system adjustments for truck arrival times and to propose a negotiation process among trucking companies and the terminal, has been proposed. In Phan and Kin (2016) the authors suggest a new appointment process by which trucking companies and terminals

collaboratively determine truck operation schedules and truck arrival appointments.

Some papers concerning different congestion issue are, among others, Sharif et al. 2011 which analyzes the potential benefits of providing real-time gate congestion information; Ambrosino and Caballini 2015 that addresses the problem of minimizing the trucks' service times at container terminals while respecting certain levels of congestion. The terminal road cycle is described in detail and a spreadsheet is used for deciding, for each truck having executed the check-in, if it should be allowed to enter the terminal and, if yes, which service level it will be given.

In this paper we investigate the management of truck arrivals by proposing a non-mandatory TAS.

Inspired by the work of Zehedner and Feillet 2013, we propose a multi-commodity network flow model for representing a general terminal where the resources are dedicated to each modal transport, and trucks approaching the terminal can decide to book or not their arrivals. A mixed integer linear programming model is proposed for determining the number of appointments to offer for each time window in such a way to serve trucks in the shortest time as possible, thus granting trucks a "good" service level.

The paper is organized as follows: in the next section we present the network flow model and the mixed integer linear programming model. Then, experimental tests on random generated data derived by a real case study of an Italian terminal are reported. Finally, conclusions and further research are given.

THE NETWORK FLOW MODEL

Starting from the model in Zehedner e Feillet 2013, we propose the following network flow for representing a general terminal. The time horizon under investigation is T and is split into s periods of time, i.e. $T = \{t_1, t_2, \dots, t_s\}$. Each truck approaching the terminal in the considered time horizon is a unit of flow that enters the network and has to exit within a given due date.

For example T can be a working day (from 6 a.m. to 10 p.m) split into 16 periods of one hour; these time periods represent the 16 time windows in which a truck can book the access to the terminal, of course, in accordance with its preferred arrival time.

Let be $C = \{1, 2, \dots, n\}$ the set of n trucks that have to approach the terminal to deliver and /or to pick up containers during the time horizon T .

Let us introduce the following notation used to characterize each truck $k \in C$:

- $p_k \forall k \in C$ represents the number of operations that truck k has to execute inside the terminal (i.e. $p_k=2$ if truck k has to deliver an export container and to pick up one import container). The maximum value of p_k

is 4 (i.e. 2 export 20' containers to deliver and 2 import 20' containers to pick up).

- r_k such that $t_1 \leq r_k \leq t_s, \forall k \in C$, represents the slot of time in which truck k approaches the terminal;
- d_k such that $r_k \leq d_k \leq t_s, \forall k \in C$, is the due time for vehicle k ; this means that truck k has to leave the terminal, having executed all p_k operations, within slot d_k .

Let us define the graph $G = (V, A)$ used for representing the terminal road cycle as described in Ambrosino and Caballini 2015 i.e. truck arrivals, the gate queue, the truck service inside the terminal and, finally, the truck exits.

V is the set of vertices given by the union of the following subsets $V = O \cup G \cup Y \cup S$ where:

$O = \{O_k \mid k \in C\}$, where O_k represents the origin node for truck k ;

$G = \{G_t \mid t \in T\}$, where G_t represents the time period t in which the trucks are at the gate queue;

$Y = \{Y_t \mid t \in T\}$ where Y_t represents the time period t in which the trucks are inside the terminal for executing the unloading/loading operations;

$S = \{S_k \mid k \in C\}$ where S_k represents the sink node for truck k .

A is the set of arcs given by the union of the following subsets $A = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5$ where:

$A_1 = \{(O_k, G_t) \mid k \in C, t = r_k\}$: for each truck k there is a link between its origin node (O_k) and the gate queue node, related to the period of time t equal to the truck arrival time (r_k).

$A_2 = \{(G_t, G_{t+1}) \mid 1 \leq t \leq t_{s-1}\}$ is the set of arcs connecting each gate queue node G_t with node G_{t+1} ; these arcs are used for representing the trucks remaining in the gate queue from time period t to $t+1$;

$A_3 = \{(G_t, Y_t) \mid t \in T\}$ for each time period t there is an arc connecting each node G_t with the corresponding node Y_t ; these arcs are used for representing trucks that in time period t enter into the terminal through the gate;

$A_4 = \{(Y_t, Y_{t+1}) \mid 1 \leq t \leq t_{s-1}\}$ is the set of arcs connecting each node Y_t with node Y_{t+1} ; these arcs are used for representing the trucks remaining inside the terminal for completing the unloading/loading operations from time period t to $t+1$;

$A_5 = \{(Y_t, S_k) \mid k \in C, r_k \leq t \leq d_k\}$ each truck k can be inside the terminal in time period t such that $r_k \leq t \leq d_k$, thus there are some links connecting nodes Y_t , of these time periods to the sink node S_k ; these arcs are used for representing the truck exit.

Each truck k is here considered as a unit of flow, that has to go from the origin node O_k to the node G_t with $t = r_k$. After that there are two possibilities: 1) the truck k is served by the gate and enters the terminal, that is the flow enters into the corresponding node Y_t ; 2) the truck k remains in the gate queue, that is the flow enters into node G_{t+1} . When the truck is inside the terminal, it has to pass from a node Y_t to the next Y_{t+1} until it has finished all

terminal operations (p_k); after that, it will leave the terminal reaching node S_k .

In Figure 1 is depicted a network flow for representing three trucks arriving at the terminal respectively in $r_1=1, r_2=2$, and $r_3=2$, having as due time $d_1=3, d_2=4$ and $d_3=3$.

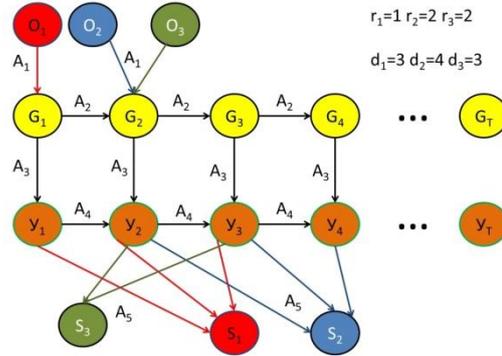


Figure 1: the network flow for 3 trucks

The Mixed Integer Linear Programming Model

The main aim of this work is to use the network flow in order to determine the best way for serving trucks that enter the terminal, i.e. to minimize the truck service time given by the total time spent by each truck inside the terminal and the total time spent in the gate queue.

Let us introduce the notation useful for presenting the mixed integer linear programming (MILP) model here proposed.

Let be:

- $a_{ij}^k \in \{0,1\}, \forall k \in C, (i,j) \in A_1$ such that $\{i=O_k, j=G_t \mid t=r_k\}$: $a_{ij}^k=1$ if truck k reaches the terminal and goes to the queue gate.
- $q_{ij}^k \in \{0,1\}, \forall k \in C, (i,j) \in A_2$ such that $\{i=G_t \mid r_k \leq t \leq d_{k-1}, j=G_{t+1} \mid r_k \leq t \leq d_{k-1}\}$: $q_{ij}^k=1$ if the unit of flow (the truck k) passes through the arc $(i,j) \in A_2$, that is the truck k remains in the queue from time period t until time period $(t+1)$.
- $e_{ij}^k \in \{0,1\}, \forall k \in C, (i,j) \in A_3$ such that $\{i=G_t \mid r_k \leq t \leq d_k, j=Y_t \mid r_k \leq t \leq d_k\}$: $e_{ij}^k=1$ if the unit of flow (the truck k) passes through the arc $(i,j) \in A_3$ that is the truck k at time period t enters the terminal.
- $y_{ij}^k \in \{0,1\}, \forall k \in C, (i,j) \in A_4$ such that $\{i=Y_t \mid r_k \leq t \leq d_{k-1}, j=Y_{t+1} \mid r_k \leq t \leq d_{k-1}\}$: $y_{ij}^k=1$ if the unit of flow (the truck k) passes through the arc $(i,j) \in A_4$, that is the truck k remains inside the terminal from time period t until time period $(t+1)$ for concluding unloading/loading operations.
- $x_{ij}^k \in \{0,1\}, \forall k \in C, (i,j) \in A_5$ such that $\{i=Y_t \mid r_k \leq t \leq d_k, j=S_k\}$: $x_{ij}^k=1$ if the unit of flow (the truck k) passes through the arc $(i,j) \in A_5$, that is the truck k in time period t has completed its p_k operations and leaves the terminal. The unit of flow k reaches its sink node.
- $b_k \in \{0,1\}, \forall k \in C$, is equal to 1 if vehicle k has a booking.

- $g_t \geq 0, g_t \in \mathbb{N}, \forall t \in T$, is the set of integer variables representing the number of gate lanes that are devoted to trucks having a booked appointment for time period t . Note that this set of variables is useful for defining the size of the booking system.
- $z_k \in \{0,1\}, \forall k \in C$ is the set of auxiliary variables used for linearizing a set of constraints in the MILP model: $z_k = \alpha_{ij}^k * b_k$.

Let us now introduce the parameter used in the model:

- $c1_{ij}^k, \forall k \in C, (i,j) \in A_2$ such that $\{i = G_t | r_k \leq t \leq d_{k-1}, j = G_{t+1} | r_k \leq t \leq d_{k-1}\}$: is the cost for having truck k in the gate queue from time period t to $t+1$;
- $c2_{ij}^k, \forall k \in C, (i,j) \in A_5$ such that $\{i = Y_t | r_k \leq t \leq d_k, j = S_k\}$: is the cost associated to truck k that leaves the terminal. The main aim is to serve each vehicle in the lower time as possible. The lower is the difference between the time period t in which the truck leaves the terminal and its due time d_k , the higher is this cost;
- α_k represents the benefit for having truck k booked;
- l_b^U is the maximum number of gate lanes that can be reserved to booked vehicles;
- π is the maximum number of vehicles that can be inside the terminal;
- π^l_t is the maximum number of unloading and loading operations that can be executed during time period t ;
- μ_b is the target service time for booked trucks fixed by the terminal management, i.e. average time that booked trucks may spend in the terminal. This average time is computed as time in the gate queue and service time for entering the terminal.
- λ represents the number of vehicles that can be processed by each gate lane in each time period.

The resulting model is the following:

M1)

$$\begin{aligned} \text{MIN} \quad & \sum_{k \in C} \sum_{(G_t, G_{t+1}) \in A_2} c1_{G_t, G_{t+1}}^k * q_{G_t, G_{t+1}}^k + \\ & + \sum_{k \in C} \sum_{(Y_t, S_k) \in A_5} c2_{Y_t, S_k}^k * x_{Y_t, S_k}^k - \sum_{k \in C} \alpha_k * b_k \end{aligned} \quad (1)$$

$$\text{s.t.} \quad a_{O_k, G_t}^k = 1 \quad \forall k \in C, t = r_k \quad (2)$$

$$\sum_{(Y_t, S_k) \in A_5} x_{Y_t, S_k}^k = 1 \quad \forall k \in C \quad (3)$$

$$(a_{O_k, G_t}^k + q_{G_{t-1}, G_t}^k) = (q_{G_t, G_{t+1}}^k + e_{G_t, Y_t}^k) \quad \forall k \in C, t | r_k \leq t \leq d_k \quad (4)$$

$$(e_{G_t, Y_t}^k + y_{Y_{t-1}, Y_t}^k) = (y_{Y_t, Y_{t+1}}^k + x_{Y_t, S_k}^k) \quad \forall k \in C, t | r_k \leq t \leq d_k \quad (5)$$

$$z_k \leq a_{O_k, G_t}^k \quad \forall k \in C, t = r_k \quad (6)$$

$$z_k \leq b_k \quad \forall k \in C \quad (7)$$

$$z_k \geq a_{O_k, G_t}^k + b_k - 1 \quad \forall k \in C, t = r_k \quad (8)$$

$$z_k \leq e_{G_t, Y_t}^k \quad \forall k \in C, t = r_k \quad (9)$$

$$\mu_b \geq \frac{1}{g_t * \lambda - \sum_{k \in C} z_k} \quad \forall t \quad (10)$$

$$g_t \leq l_b^U \quad \forall t \quad (11)$$

$$\sum_{k \in C} y_{Y_t, Y_{t+1}}^k \leq \pi \quad \forall t | t \leq |T|-1 \quad (12)$$

$$\sum_{k \in C} p_k * x_{Y_t, S_k}^k \leq \pi_t^p \quad \forall t \quad (13)$$

Objective function (1) minimizes the costs associated to the trucks in the queue at the gate and the costs for serving the truck near their due time, while trying to maximize the total number of booked vehicles.

Constraints (2) force each vehicle k to arrive at its preferred time r_k , i.e. each truck k leaves its source node in r_k , while constraints (3) ensure to each vehicle k to reach its sink node.

(4) and (5) are the flow conservation constraints for nodes of set G and Y , respectively.

Constraints (6), (7), (8) define variable z_k as the product of variables a_{ij}^k and b_k in order to preserve linearity in the model. These constraints, together with constraints (9), replace the following ones that are necessary to guarantee that a booked truck k enters the terminal as soon as it arrives at the gate:

$$a_{O_k, G_t}^k * b_k \leq e_{G_t, Y_t}^k \quad \forall k \in C, t = r_k$$

Constraints (10) limit the time spent in the queue outside the terminal by booked vehicles, this maximum queue time is decided by the terminal operator and it is set by parameter μ_b .

Set of constraints (11) bounds the number of gate lanes reserved to booked vehicles, while constraints (12) limit the number of vehicles inside the terminal from time period t to the following $t+1$.

Lastly, constraints (13) represent the terminal capacity for handling operations: the total number of operations executed in each time period t must be no greater than the terminal handling capacity.

An Extension of the Previous Network Flow Model

A drawback of model M1 is the following: a truck can not book the access to the terminal in a time period different from its preferred arrival time r_k . Anyway, for increasing the capability of the booking system of modifying the arrival process to the terminal it is necessary to give the possibility to the appointment system to book for period of times different from preferred truck arrivals.

The disadvantage for the truck operator that is obliged to modify his behavior, will be compensated by the advantages derived by having fixed and known the service time at the terminal. Moreover, the allowable deviation between preferred and booked time period is limited.

Thus, we have modified both the network flow and the model in such a way to permit to a truck k to book an appointment for a time period different from r_k .

In Figure 2 the new network flow is depicted.

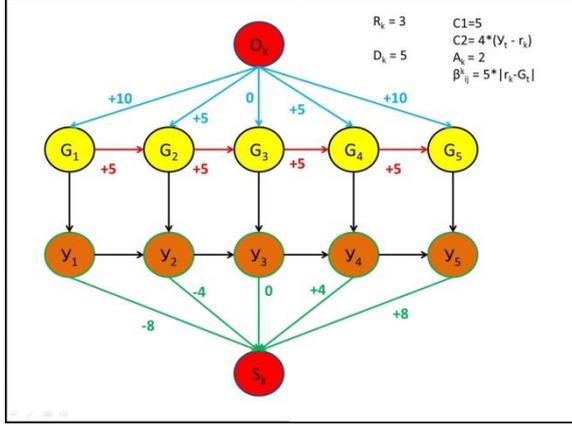


Figure 2: example of network flow for one truck with a maximum absolute deviation of 2 periods

Let be Δ the maximum deviation from the preferred arrival time (r_k) of vehicle k , thus follows:

- $A_1 = \{(O_k, G_t) \mid k \in C, r_k - \Delta \leq t \leq r_k + \Delta\}$;
- $A_5 = \{(Y_t, S_k) \mid k \in C, r_k - \Delta \leq t \leq d_k\}$;
- $a^{k_{ij}} \in \{0,1\}, \forall k \in C, (i,j) \in A_1$ such that $\{i = O_k, j = G_t \mid r_k - \Delta \leq t \leq r_k + \Delta\}$;
- $x^{k_{ij}} \in \{0,1\}, \forall k \in C, (i,j) \in A_5$ such that $\{i = Y_t \mid r_k - \Delta \leq t \leq d_k, j = S_k\}$;
- $q^{k_{ij}} \in \{0,1\}, \forall k \in C, (i,j) \in A_2$ such that $\{i = G_t \mid r_k - \Delta \leq t \leq d_{k-1}, j = G_{t+1} \mid r_k - \Delta \leq t \leq d_{k-1}\}$;
- $e^{k_{ij}} \in \{0,1\}, \forall k \in C, (i,j) \in A_3$ such that $\{i = G_t \mid r_k - \Delta \leq t \leq d_k, j = Y_t \mid r_k - \Delta \leq t \leq d_k\}$;
- $y^{k_{ij}} \in \{0,1\}, \forall k \in C, (i,j) \in A_4$ such that $\{i = Y_t \mid r_k - \Delta \leq t \leq d_{k-1}, j = Y_{t+1} \mid r_k - \Delta \leq t \leq d_{k-1}\}$;
- z_k is changed as z^k_i such that $k \in C$ and $i = G_t \mid r_k - \Delta \leq t \leq d_k$, to better explain the product $a^{k_{ij}} * b^k$ including the time of arrival at the gate.

Constraints of model M1) defined for t such that $r_k \leq t \leq d_k$, must now be defined for $r_k - \Delta \leq t \leq d_k$, and those defined for $t = r_k$ must now be defined for $r_k - \Delta \leq t \leq r_k + \Delta$. Moreover, constraints 2) must be modified as follows:

$$\sum_{(O_k, G_t) \in A_1} a^{k_{O_k, G_t}} = 1 \quad \forall k \in C$$

Finally, in the new model (from now we will refer to it as M2) we try to minimize also the deviation between the trucks appointments and their preferred arrival time (r_k). For this reason we define:

- β the cost paid for having modified the arrival time at the terminal.

The new component of the objective function is:

$$\beta * \sum_{k \in C} \sum_{i = G_t} (|t * z_i^k - r_k|)$$

COMPUTATIONAL RESULTS

In this section we report the results obtained by solving model M2 for some random generated data derived by a real case study of an Italian terminal.

Figure 3 shows the arrival distribution in a working day of the terminal used for generating data. As usual this distribution presents a peak of the truck arrivals.

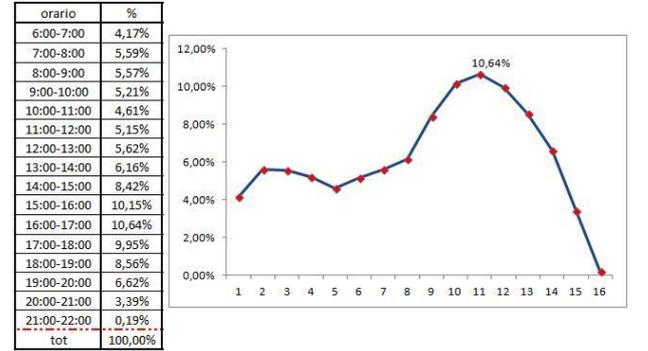


Figure 3: arrival distribution

In this experimental campaign, we have considered a wide range of different scenarios.

First of all, we have analyzed the differences between implement a 1-hour time window and a 2-hours time window, where the latter one offers an easier organization for haulers, at the cost of less efficiency for the terminal.

Secondly, we have simulated three different possibilities for the daily total arrival of trucks: in the first case the total number of vehicles results less than the average of the work situation (1.800 trucks per day), in the second one the total number of vehicles results in average with the work situation (2.000 trucks per day) and in the last scenarios the total number of vehicles results higher than the average work situation (2.200 trucks per day).

To deeply analyze the behavior of the model, and mostly to validate it, we have also generated scenarios by changing the number of operations to be executed at the terminal by trucks.

Moreover, to study how the policy of assured maximum time elapsed in the queue impacts on the model solutions, various values of μ_b have been applied.

Lastly, we have used three different average times for gates operations (λ) to identify the operational target terminal handlers should aim for.

Each scenario has been named following the nomenclature shown in the Table below.

		Value			
		1	2	3	4
A	Time windows width	1 hour	2 hours		
B	Total daily arrivals	1800	2000	2200	
C	Number of operations (partition)	50/50/0/0	40/40/20/0	40/40/15/5	
D	Maximum time in queue	10 minutes	20 minutes	30 minutes	40 minutes
E	Handling capacity (number of vehicles served)	60	90	120	

Table 1: Scenarios nomenclature

All generated scenarios have been solved up to optimality by model M2.

By varying μ_b , we can observe in Figure 4 little changes in the optimal arrival distributions.

Solutions obtained by imposing a maximum time to spend in the gate queue for booked trucks of 40 minutes do not differ much from those obtained by having this limit equal to 10 minutes. Thus terminal operator can persuade hauler to book their entry by promising a smaller amount of time spent in the queue while limiting the risks of operational problems.

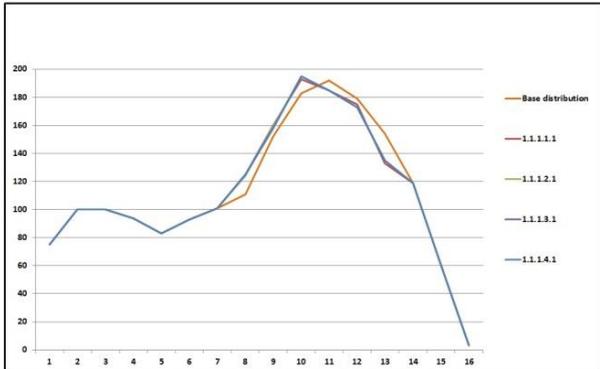


Figure 4: arrival distribution varying μ_b

More interesting is the study of the solutions obtained by varying parameters λ . As shown in Figure 5 when the time required for serving one truck at the gate passes from 60 to 45 seconds, a significant redistribution of trucks arrivals can be observed, while passing from 45 to 30 seconds significant changes are no more evident.

This is due to the fact that in these scenarios, the number of entrances is limited by terminal handling capacity, where in the case with 45 seconds both constraints, time spent in queue and handling capacity, affect the solution.

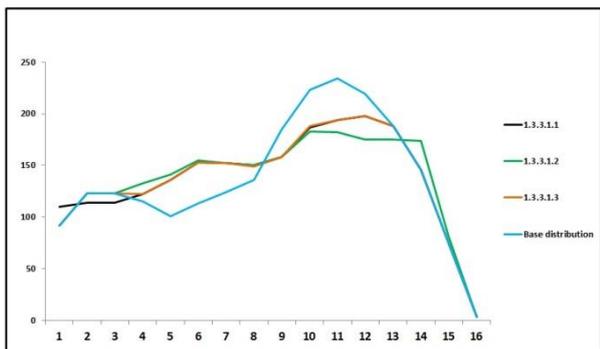


Figure 5: arrival distribution varying λ

Increasing the number of operations to be executed at the terminal by trucks, pushes the model to increasingly modify arrivals distribution.

A lower number of operations associated to each truck allows the terminal to easily handle the whole traffic, especially with a low number of daily arrivals, while the terminal reaches its full handling capacity when an higher

number of operations is considered. The graph of Figure 6 shows how the model modifies the arrival distributions.

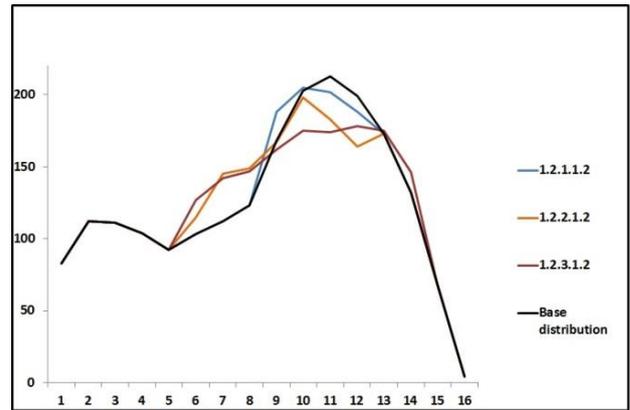


Figure 6: arrival distribution varying the number of operation to execute

Similarly, the model increases its action in redistributing trucks arrivals at terminal when the daily total number of trucks ($|C|$) increases, as shown in Figure 7.

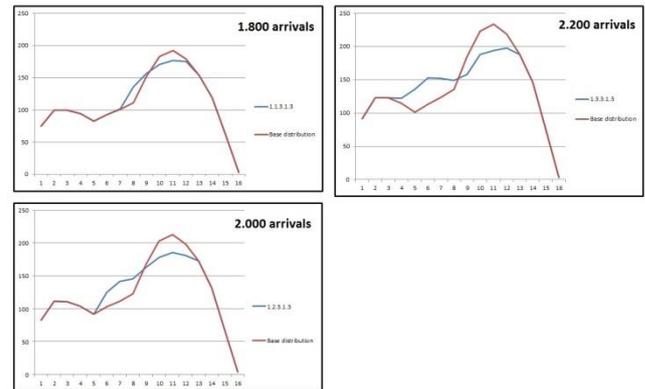


Figure 7: arrival distributions for different and increasing $|C|$

CONCLUSIONS AND FUTURE WORKS

In order to tackle the problem of increasing queues outside container terminals, the implementation of a non-mandatory truck appointment system is studied.

Optimal size of the system, capable of decreasing congestion while trying to impact as little as possible hauler operational plans, is obtained by a mixed integer linear optimization model, used for defining optimal flows in the network flow model which simulates the operability of a container terminal during a workday.

Solutions found demonstrate the effectiveness of the model to re-shape arrivals distribution and to adapt to shifting conditions, both external (number of arrivals, number of operations to be executed inside the terminal) and internal (mostly operational target, such as maximum

time spent in the queue by booked vehicles and average service time at gates).

Future developments focus on the modification of the network flow, in such a way to better simulate the behavior of trucks once inside the perimeter of a container terminal. For this reason the width of the time windows will be decreased. Moreover, a limitation on the time spent in queue for not booked vehicles will be included in the model.

The optimization model will be developed using C# programming language, this addition allows the simulation of a wider number of scenarios for both a better understanding of the problem and validating the model.

REFERENCES

- Alessandri, C.; M. Cervellera; M. Cuneo; and M. Gaggero. 2008. "Nonlinear Predictive Control for the Management of Container Flows in Maritime Intermodal Terminals." *Proceedings of the 47th IEEE Conference on Decision and Control Cancun, Mexico, Dec. 9-11.* .
- Ambrosino, D. and C. Caballini. 2015. "Congestion and truck service time minimization in a container terminal." *Maritime-Port Technology and Development. Proceedings of the International Conference on Maritime and Port Technology and Development, MTEC 2014, 1-10.*
- Ambrosino, D. and A. Sciomachen. 2014. "Location of Mid-Range Dry Ports in Multimodal Logistic Networks." In: Raffaele Cerulli, Giovanni Felici, Anna Sciomachen. *Procedia - Social and Behavioral Sciences.* Elsevier, No 108, 118-128.
- Chen, G.; K. Govindan; and Z. Yang. 2013. "Managing truck arrivals with time windows to alleviate gate congestion at container terminals." *International Journal of Production Economics*, No 141, 179–188.
- Chen, G.; K. Govindan; and M.M. Golias. 2013b. "Reducing truck emissions at container terminals in a low carbon economy: proposal of a queueing-based bi-objective model for optimizing truck arrival pattern." *Transportation Research Part E*, No 55, 3–22.
- Chen, X.; X. Zhou; and G.F. List. 2011. "Using time-varying tolls to optimize truck arrivals at ports." *Transportation Research Part E*, No 47, 965–982.
- Cullinane, K. and G. Wilmsmeier. 2011. "The contribution of the dry port concept to the extension of port life cycles." In: Bose, J.W. (Ed.), *Handbook of Terminal Planning.* Springer, New York, 359–379.
- Dekker, R.; S. van der Heide; E. van Asperen; and P. Ypsilantis. 2013. "A chassis exchange terminal to reduce truck congestion at container terminals." *Flexible Service Manufacturing Journal*, No 25, 528–542.
- Guan, C.Q. and R.F. Liu. 2009. "Container terminal appointment system optimization." *Maritime Economics and Logistics*, No 11(4), 378–398.
- Huynh, N. and C. M. Walton. 2008. "Robust Scheduling of Truck Arrivals at Marine Container Terminals." *Journal of Transportation Engineering.* August, 347-353.

- Morais, P. and E. Lord. 2006. "Terminal appointment system study." *Transport Canada, Ottawa.* <<http://trid.trb.org/view.aspx?id=792769>>.
- Phan, T.M.H. and K.H. Kim. 2015. "Negotiating truck arrival times among trucking companies and a container terminal." *Transportation Research Part E*, No 75, 132–144.
- Phan, T.M.H. and K.H. Kim. 2016. "Collaborative truck scheduling and appointments for trucking companies and container terminals." *Transportation Research Part B*, No 86, 37–50.
- Roso, V.; J. Woxenius; and K. Lumsden. 2009. "The dry port concept: Connecting container seaports with the hinterland." *Journal of Transport Geography*, No 17-5, 381–398.
- Sharif, O.; N. Huynh; and J.M. Vidal. 2011. "Application of El Farol model for managing marine terminal gate congestion." *Research Transport Economics*, No 32, 81–89.
- Zehndner, E. and D. Feillet. 2014. "Benefits of a truck appointment system on the service quality of inland transport modes at a multimodal container terminal." *European Journal of Operational Research*, No 235, 461–469.
- Zhang, X.; Q. Zeng; and W. Chen. 2013. "Optimization model for truck appointment in container terminals." *Procedia-Social and Behavioral Sciences* No 96, 1938–1947.

AUTHOR BIOGRAPHIES



DANIELA AMBROSINO is assistant professor of Operation Research at the University of Genoa, Italy. She is teaching Operation Research for management and Optimization and Simulation for maritime transport.

Her main research activities are in the field of distribution network design and maritime logistic. Her e-mail address is: daniela.ambrosino@economia.unige.it



LORENZO PEIRANO is postgraduate in Maritime Economics and Ports Management at the University of Genoa, Italy. He occasionally collaborates with the University of Genoa for research activities in maritime logistics. He is working at Aldieri Autotrasporti S.p.A.

(Italy) as airfreight forwarder.

This paper is mainly inspired by his postgraduate degree dissertation awarded by the Italian society of Operational Research in September 2015.

His email address is lorenzo.peirano@gmail.com

AN APPLICATION OF DISCRETE EVENT SIMULATION ON ORDER PICKING STRATEGIES: A CASE STUDY OF FOOTWEAR WAREHOUSES

Thananya Wasusri and Prasit Theerawongsathon
Logistics Management Program, Graduate School of Management and Innovation
King Mongkut's University of Technology Thonburi
126 Prachautid Rd., Bangkok, Thailand
E-mail: Thananya.was@kmutt.ac.th

KEYWORDS

Order Picking, Batch Picking, Zone Picking, Simulation.

ABSTRACT

A footwear business is one of highly competitive markets. Footwear businesses must continuously improve three main key performance measures namely quality, cost and lead time. The case study is one of footwear businesses in Thailand. As customer demands are fragmented, the company needs to offer different products to serve customers' satisfaction. Customer orders are then small quantities, but contain several different product types. As a result, picking activities in its warehouse are becoming more difficult especially for sport shoes, fashion shoes and sandals. While individual order picking strategy was utilized, it caused long lead time in its picking process. Its delivery performance cannot be achieved. Discrete event simulation modeling using ARENA was conducted to investigate the performance of one order picking, batch picking and zone picking. It was found that the best alternative is zone picking with 4 orders per batch. The picking time can be shortened about 15%.

INTRODUCTION

It has been recognized that quality, cost and time are major key performance indicators for almost all companies need to keep tracking and improving. Shortening throughput and delivery lead times can offer a competitive advantage as quick responses to customer demand uncertainties or requirements are inevitable. Warehouses provide an important link in supply chains, where products can be temporarily stored and retrieving products from storage can be managed regarding customer orders (Petersen 2002). To manage warehouses, there are main warehouse activities to consider such as receiving, put-away, storage, order picking, packing, loading stock counting, value-adding services (Richards 2014). The cost of order picking is estimated to be about 55% of the total warehouse cost as order picking has been considered as the most labour-intensive and costly activity which almost all warehouses (De Koster et al. 2007). Order picking is then a retrieving process of products or items from warehouse storage locations to fulfil customer orders

(Petersen et al. 2004). To improve the efficiency of order picking process can shorten supply chain lead time and reduce warehousing and supply chain costs.

To improve order picking efficiency, order picking strategies are being utilized. Discrete picking or individual picking is where a picker is assigned to pick all the items in a single order for a pick-tour. In batch picking, many orders are grouped or batched together and a picker picks all the items for a given batch. Zone picking assigns each picker to a specific zone or zones and is responsible for picking the items in that zones (Parikh and Meller 2008). Warehouses should select and apply those strategies to fit with their nature of products and demand patterns.

In view of order picking process, a footwear company in Thailand has explored its order picking process. The objective of this study is to investigate possibilities of applying picking strategies to its order picking process and evaluate its benefits based on time and labor utilization. The next section provides the background of the case study, followed by a survey of the related literature. Research methodology and results are proposed. Finally, conclusions and suggestions are described.

BACKGROUND OF THE CASE STUDY

The case study, called "ABC", is a Thai footwear company and has its own brand products. As customer demands are fragmented and fast-changing, ABC has offered a variety of product to customers that are school shoes, fashion shoes, sport shoes and sandals. To serve its customers, ABC has its own warehouse to store and replenish the products to their customers that are modern trades, wholesalers, retailers and also e-commerce customers. The warehouse is equipped with racks, forklifts, barcoding, wifi, warehouse management system (WMS), and safety systems.

Warehouse Activity

The warehouse activities include purchase order entry, goods receiving, put-away, picking, invoicing, checking, dispatch, reverse logistics, replenishment and stock count. Purchase order entry is dealing with supply planning and purchase order placing to its factory. Goods receiving process starts from preparing goods

receipt schedule, preparing storage space, checking products received against the goods receipt schedule and record in WMS. Put-away activity is to move the products received to the locations assigned with using barcoding and handhelds. Picking process includes picking list formulation, picking assignment and picking travelling and moving to marshalling lanes. According to the limitation of WMS, the only picking strategy available is individual order picking.

After products are picked, labelling the products and issuing invoices are the next main activities. Checking process is conducted to insure that the products picked and invoice are correct. Dispatching process is to move the products into trucks or vans and passing the ownership to third-party logistics. Reverse logistics is to received returned products from customers and invoice deduction. Replenishment is to move products from reserved area to picking face area. Stock counting is to count all products in the warehouse. It was found that inventory accuracy of the warehouse is 97%. In the warehouse, there are eleven staff. Five staffs are mainly working to manage all information and planning such as purchasing planning, invoice issuing and other documenting processing. Six staffs are assigned to be pickers.

Warehouse layout

The warehouse area is 4,800 square meters and consists of ten racks, namely A to J. Each rack contains five levels and the ground floor level is assigned to be pick face areas. From the second floor level to the fifth floor level are specified as reserved areas. The put-away process is performed by random location generated by WMS.

Demand characteristics

There are 13,355 active stock keeping units (SKUs) in the warehouse. About 80% SKUs is school shoes which has its 2 high seasons during summer vacation and October vacation. For two high seasons, the movement of products are almost full-pallet picking. Individual order picking is appropriate to handle during high season periods. During off-peak season, customer orders are small quantities, but several different product types. The effect of individual picking process caused fatigue in pickers as they had to walk very long distance. It was difficult to balance work-load. In addition, rush orders cannot be finished on time. It was then needed to improve picking process in off-peak season.

LITERATURE REVIEW

Richards (2014) classified order picking strategies three categories that are picker to goods, goods to picker and automated picking. The majority of warehouses continue operating with picker-to-goods operations. Picker to goods strategy consists of individual order picking, cluster picking, batch picking, zone picking and wave picking. Individual order picking or discrete

order picking is a picker takes one order and travels through the warehouse on foot or using forklifts to collect items until the whole order is completed. Peterson and Aase (2004) also noted that discrete order picking is often preferable because it is easy to implement and order integrity is always maintained.

Cluster picking, a picker can have many orders at a time and travels around the warehouse to pick items into individual compartments on their trolleys or pick carts. In other words, picking and sorting can be done at the same time by using trolleys or pick carts that provide individual compartments. Cluster picking can extend to use with pick-to-light technology as well (Richards 2014).

Batch picking is similar to cluster picking as operators pick many orders at the same time, but orders are consolidated or batched into a picking list. After picking all items will be sorted or allocated into each customer order. The batch picking is then pick-and-sort (Richards 2014). Peterson and Aase (2004) described batch picking as combining several orders into batches. First-come-first-served (FCFS) batching can combine or consolidate orders as they arrive until the maximum batch size has been reached. This can be a way that order batching can be conducted. Parikh and Meller (2008) provided insights of batch picking. Batch picking can be classified into two categories that are pick-and-sort and sort-while-picked. Batch picking with pick-and-sort method can increase pick-rate of pickers because sorting is not a part of picking and decrease chances of workload-imbalance. At the same time, probability of blocking can increase as a result, pick tours can be long. Batch picking with sort-while-pick method can decrease chances of workload-imbalance and does not require a sorting system, while pick-rate can decrease as sorting is a part of picking process. In addition, probability of blocking can be increased.

Zone picking is where picking is defined by areas in the warehouse. Pickers are assigned to pick from a particular zone or zones. Orders can be picked at the same time within the zones and the items picked will be sorted or allocated according to each customer order later (Richards 2014). Parikh and Meller (2008) described zone picking with sequential or progressive method is to pick one order at a time and in one zone at a time. In contrast, zone picking with simultaneous or synchronized method is to pick all items regarding batched orders are picked at the same time from all the zones and then orders are sorted. Zone picking with progressive can increase a pick-rate of pickers as pick-tours are short and do not need a sort system. It can eliminate blocking problems, but it can increase chances of workload-imbalance. Zone picking with synchronized method can increase a pick-rate of pickers as pick-tours are short and eliminate blocking as well, but it requires a

sorting system and may result in increasing workload-imbalance.

Wave picking is that orders are consolidated and released at particular times to associate them with vehicle departures or replenishment cycles (Richards 2014). Peterson and Aase (2004) noted that wave picking is the combination of batching and zoning into “wave” picking where a picker is responsible for SKUs in their zone for several orders. The benefit of these policies become evidently when the size of the warehouse increases, but zone picking needs sorting operations to consolidate orders from the different zones. Marchet et al. (2011) noted that the pick-and-sort system is to pick a number of each single item from the batching of multiple orders (wave) and place them at the sorting area. As a result of order accumulations, the same item will be picked at the same time. It can reduce the number of different locations in pick-tours or so called ‘overlapping effect’. The pick-and-sort system is then based on picking waves. The period of time in which a groups of orders is picked in one picking area before moving to the next area such as sorting area.

To improve order picking process, companies require to apply order picking strategies and techniques that are appropriate to the nature of product, the size of order and the quantity of items. Studies have been conducted to improve order picking process. Tang and Chew (1997) conducted a simulation study to review the effect of batch size on tardiness. They noted that smaller batch size could reduce the average delay time when compared to one order picking with unlimited resources. Petersen (2002) studied the effects of picking zone configuration on picker distances. The storage size of the picking zone, picking policies and the number items of pick list were included. As an example, batch zone operations normally have large pick lists per zone and would perform well with zone configuration having one or two aisles. However, volume-based storage affects less travel distances than random storage for all zone configurations. Peterson and Aase (2004) conducted intensive simulation modelling to investigate several picking, storage, and routing policies. Several sensitivity analyses are constructed to evaluate the effect of order size, warehouse shape, location of pick-up/drop-off point, and demand distribution on performance. Batching of orders can offer the greatest savings especially when smaller order sizes are usual. Rim and Park (2008) proposed a linear programming model to conduct border batching with subject to minimize the order fill rate. Then, a simulation study was conducted to investigate the performance of LP and FCFS on order batching. Although the FCFS rule has an advantage of smoothing the picking load and reducing the number of pickers, the proposed LP performs better in terms of the order fill rate by 12.7%. It was pointed out that the order fill rate creases when the number of orders or items increases. Planning order picking can

increase the order fill rate. As order batching methods can significantly affect the order fill rate and customer service levels, Henn and Schmid (2013) applied metaheuristics in order batching and sequencing. The proposed metaheuristics can improve the total tardiness of a given set of customer orders by 46%.

To improve and measure picking process, not only picking efficiency or productivity have been investigated, but the cost model has been proposed. Parikh and Meller (2008) developed a cost model to estimate the cost of batch picking and zone picking. The cost model included the cost of pickers, equipment, imbalance, sorting system, and packers. It was found that workload-imbalance is greater in zone picking comparing to batch picking. workload-imbalance is more important when the order sizes increase, item distribution is not likely to be uniform, and the number of waves increases. Marchet et al. (2011) introduced an analytical model to estimate the picking efficiency regarding wavelength. Simulation modelling has been conducted to validate the analytical model. A case study of book distributors was presented to describe how the model can be applied to estimate picking efficiency based on the size of the wave and to determine the trade-off between picking efficiency and sorting cost. The results show that the number of waves has a significant effect on the pick-and-sort system. The number of picking waves per day should not be too small that is a long wavelength per pick, as it can crease the sorting costs and the picking efficiency could not be balanced. The appropriate number of waves cannot be generalizable and it must be determined based on the operating environment. The further study may propose an extension to include considerations in terms of peak versus non-peak activity levels, and a generalization through the modelling of the costs and activities of the pick-and-sort systems.

Muanul picking and automated picking has been studied by Lee et al. (2015). They conducted the experiments to evaluate the performance of manual order picking and pick-to-light picking or digital supported manual order picking that can help pickers to seek for locations and amount of the items on the storage or shelves. It was found that the pick-to-light technology can reduce both mean and standard deviation values of picking time.

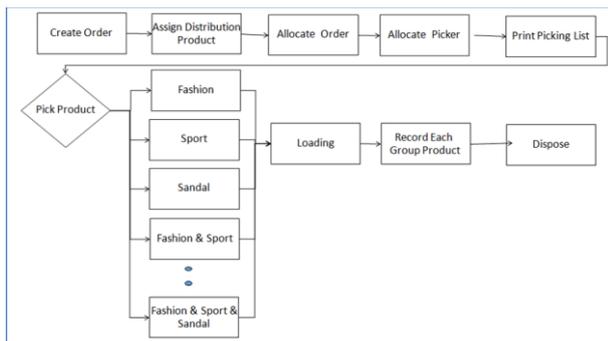
It can be seen that many authors are interested in exploring tools or techniques to improve picking processes with regarding some key performance measures such as time, tardiness, distance, workload and cost. However, the picking configurations proposed cannot be generalized for all circumstances. This research is then constructed from a real life situation and concerns both time and workload. The techniques investigated are batching and zoneing as they can be implemented in the case study without no adding too much cost comparing to automated picking systems.

Simulation modelling can be a tool to investigate the effects of order picking. The next part is research methodology used to conduct this research.

RESEARCH METHODOLOGY

Discrete event simulation, ARENA, is applied as a tool to study this research. Off-peak sales seasonal is selected to study its effect on picking process as customer orders are small and high-variety. The current picking process includes receiving customer orders, allocating orders, assigning picker, printing pick list, picking travel (individual order picking), and dispatching. From the current process, it can be written as a flow chart shown in figure 1. The results from the current situation are used as a base line to compare with other alternatives.

Batch picking and zone picking are studied based on one wave per day. FCFS is the technique applied to batch or group customer orders. Batch and zone picking proceses will start from order receiving, allocating orders (batching orders), assigning pickers based on batch picking and zone picking, printing picking list, pick tours, sorting and dispatching.



Figures 1: The Current Picking Process

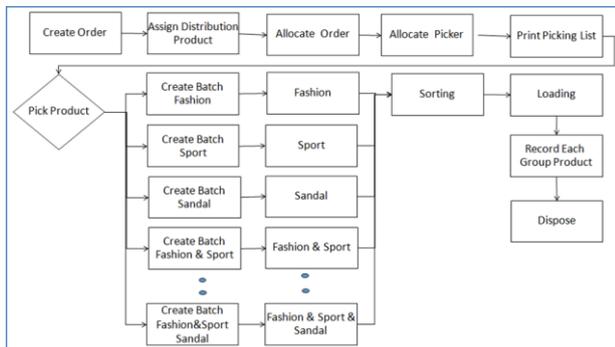


Figure 2 : Batch and Zone Picking Processes

Scenarios proposed

To investigate the effects of picking strategies, three main scenarios are proposed. The current situation, discrete order picking is investigated. Batch and zone picking are examined based on 2 to 6 orders grouping.

Two main performance measures are evaluated that are staff utilization and lead time.

Table 1 : Scenarios Proposed

Picking policy	Order grouping					Performance measures
Discrete order	none					Staff utilization and time
Batch picking	2	3	4	5	6	
Zone picking	2	3	4	5	6	

Verification and validation

Daily operations for 180 days were collected to use in the simulation program conducted. Every input datum is fitted with theoretical probability distribution by using input analyzer module in ARENA. Verification has been conducted for each module in order to assure that the simulation program can perform as required. Five key variables are receiving orders, allocating orders, printing picking list, order picking and dispatching used to validate with real life performances. It was found that the five key performances obtained by simulation are not significantly different from the actual performances as shown in table 2.

Table 2 : Output Validation

	Actual (min.)		Simulation (min.)	
	Average	Half Width	Average	Half Width
Checking orders	9	0.45	8.33	0.02
Printing picking list	1	0.05	0.999	0
Order picking				
Fashion	6.98	0.27	7.95	0.09
Sandals	18	0.9	17.13	0.23
Sport	16	0.8	15.67	0.54
Dispatching	3	0.08	3.00	0.01

To identify run length and run replication, equation 1 was used to identify number of replications (Kelton et al. 2007)

$$n \cong n_0 \frac{h_0^2}{h^2} \tag{1}$$

The expected number of replications is represented by n. The desired half width is h and n₀ represents the number of replications for the pilot run. The half width obtained from the pilot run is h₀. The desired half width picking time is less than 5 minutes. The simulation is conducted based on terminating system as the case study starts working from waiting for customer orders and finishing its work once all customer orders received have been dispatched. The number of replications is twenty.

Although batch picking and zone picking never utilized with the case study, a time and motion study was conducted to obtain some processing time variables to apply in batch picking and zone picking simulations. The next part is results from simulation runs.

RESULTS

Current situation

The simulation on current situation that is using individual picking has been conducted. It was found that the longest lead time in the process studied is allocating orders. The allocating order activity takes about 68.57 minutes per order. During allocating order, there is waiting time in process as well called 'cut off order'. The main reason of cut off order is to consolidate all orders to dispatch. For picking activities, seven types of picking has been found that are to pick within product family, combining product families. Picking three different product families can result in 37.56 minutes per order, as shown in table 3. In addition, total process time for each order type has been collected and shown in table 4. It can be seen that increasing product variety leads to increasing total process time per order. Applying individual order picking results in low staff utilizations. Table 5 shows staff utilizations and all pickers are found to be busy no more than 60% of their total time.

Batch picking

Batch picking has been explored by batching or grouping two orders per batch until six orders per batch. When batch picking is performed, total process time per order can be reduced. Picking time is shortened, but allocating order time is increased. Allocating order time increases because batching order is conducted by manual as existing WMS does not support this activity.

Table 3 : Current Situation Results

Process	Time per one order (min.)	
	Average	Half Width
Checking orders	8.33	0.02
Allocating orders	68.57	2.47
Printing picking list	0.999	0
Order picking		
Fashion	7.95	0.09
Sandals	17.13	0.23
Sport	15.60	0.54
Fashion & Sandals	23.51	1.3
Fashion & Sport	19.50	2.28
Sport & Sandals	30.44	2.56
Fashion & Sport & Sandals	37.56	3.18
Dispatching	3.00	0.01

Moreover, increasing number of batching will directly increase waiting time to group or batch customer orders. Only a three-order-batch picking is selected to illustrate in more details. Table 6 shows total process time per order of a three-order-batch picking policy. Not only can the total time be reduced, the utilization of staff can be reduced as well, as shown in table 7.

Table 4: Current Total Process Time per Order

Types	Average (min.)	Half width (min.)
Fashion	88.86	2.59
Sandals	98.04	2.73
Sport	96.51	3.04
Fashion & Sandals	104.42	3.8
Fashion & Sport	100.41	4.78
Sport & Sandals	111.34	5.06
Fashion & Sport & Sandals	118.47	5.68

Table 5 : Current Staff Utilization

Staff	Utilization (%)	
	Average	Half width
Picker 1	53	0.01
Picker 2	54	0.01
Picker 3	57	0.01
Picker 4	52	0.01
Picker 5	55	0.01
Picker 6	50	0.01

From the experiments, it was found that a four-order batch can give the shortest process time as a six-order batch results in higher total process time due to increasing of allocating order time and sorting time, as shown in table 8. Moreover, when many orders are grouped, walking distances can be reduced and it can be shown in terms of staff utilization shown in table 9.

Table 6 : A Three-Order-Batch Picking Time per Order

Types	Average (min.)	Half width (min.)
Fashion	80.48	2.38
Sandals	86.13	2.67
Sport	103.10	5.93
Fashion & Sandals	98.23	8.01
Fashion & Sport	81.35	5.84
Sport & Sandals	79.95	5.37
Fashion & Sport & Sandals	94.93	3.28

Table 7 : A Three-Order-Batch Staff Utilization

Staff	Utilization (%)	
	Average	Half width
Picker 1	17	0.01
Picker 2	17	0.01
Picker 3	18	0.01
Picker 4	18	0.01
Picker 5	16	0.01
Picker 6	16	0.01

Table 8 : Average Total Process Time of Batch Picking

Types	Average total process time (min/order)				
	As-Is	Batch picking			
		2	3	4	6
Fashion	88.86	81.63	80.48	79.84	84.75
Sandals	98.04	88.14	86.13	84.87	89.06
Sport	96.51	99.03	103.10	101.38	100.03
Fashion & Sandals	104.42	107.76	98.23	83.10	82.33
Fashion & Sport	100.41	97.20	81.35	78.31	0.00
Sport & Sandals	111.34	92.66	79.95	76.82	0.00
Average	99.93	94.40	88.20	84.05	59.36

Table 9 : Staff Utilization of Batch Picking

Types	Utilization (%)				
	As-Is	Batch picking			
		2	3	4	6
Picker 1	53	27	17	13	9
Picker 2	54	27	17	14	9
Picker 3	57	27	18	12	8
Picker 4	52	26	18	11	6
Picker 5	55	26	16	11	6
Picker 6	50	26	16	12	7

Zone picking

Zone picking simulation has been conducted by batching or grouping two orders per batch up to six orders per batch. The results are similar to those of batch picking. When having more orders grouping, picking time can be shortened, but allocating order time and sorting time will increase. It was found that a four-order batching can offer the shortest process time as shown on table 10. However, zone picking can give a better result in term of staff utilization as three staff can be reduced. Table 11 shows utilization of staff when zone picking is applied.

CONCLUSION

From the results, it can be concluded that individual picking strategy cannot perform well due to the nature of orders that are small and contain a variety of products. Zone picking and batch picking with using FCFS strategy to group orders are found to be preferable. The both two strategies can shorten total

process time, but zone picking performs better than batch picking in terms of staff utilization. Therefore, zone picking can be appropriately implemented during low season. To get a better result, other batching techniques should be investigated as they may shorten picking time. Batch sizing has significantly affected mean order throuput time. From our investigation, batching up to four orders is recommended.

Table 10 : Average Total Process Time of Zone Picking

Types	Average total process time (min/order)				
	As-Is	Zone picking (Batch size)			
		2	3	4	6
Fashion	88.86	80.62	80.26	79.13	84.19
Sandals	98.041	86.92	85.85	84.32	88.52
Sport	96.51	101.22	102.83	102.20	99.63
Fashion & Sandals	104.42	110.25	99.72	85.35	81.63
Fashion & Sport	100.41	97.01	86.12	77.79	0.00
Sport & Sandals	111.34	97.13	81.95	76.69	0.00
Average	99.93	95.52	89.45	84.24	58.99

Table 11 : Staff Utilization of Batch Picking

Types	Utilization (%)				
	As-Is	Zone picking (Batch size)			
		2	3	4	6
Picker 1	53	27	21	17	23
Picker 2	54	6	4	2	2
Picker 3	57	26	20	15	21
Picker 4	52	0	0	0	0
Picker 5	55	0	0	0	0
Picker 6	50	0	0	0	0

Moreover, if WMS would need to modify to support zone picking, simulation modelling should be conducted to cover high season periods to investigate whether or not individual order picking can perform better than zone picking. To modify WMS, order allocating and sorting activites will be added.

REFERENCES

De Koster, R.; T. Le-Duc. and K.J. Roodbergen. 2007. "Design and control of warehouse order picking: A literature review." *European Journal of Operational Research*, No.182, 481-501.

Henn, S. and V. Schmid. 2013. "Metaheuristics for order batching and sequencing in manual order picking systems." *Computers & Industrial Engineering*, No.66, 338-351.

Kelton, W.D.; R.P. Sadowski. and N.B. Zupick. 2015, *Simulation with Arena, 6th Edition.*, McGraw-Hill, New York.

Lee, J.A.; Y.S. Changa; H.J. Shimb and S.J. Chob. 2015. "A study on the picking process time." *Procedia Manufacturing*, No. 3, 731-738.

- Marchet, G.; M. Melacini and S. Perotti. 2011. "A model for design and performance estimation of pick-and-sort order picking systems." *Journal of Manufacturing Technology*, No.2, 261-282.
- Parikh, P.J. and R. D. Meller. 2008. "Selecting between batch and zone order picking strategies in a distribution center", *Transportation Research Part E*, No. 44, 696–719.
- Petersen, C.G. 2002. "Considerations in order picking zone configuration." *International Journal of Operations & Production Management*, No.7, 793-805.
- Petersen, C.G. and G.R. Aase. 2004. "A Comparison of picking, storage, and routing policies in manual order picking." *International Journal of Production Economics*, No. 92, 11-19.
- Petersen, C.G.; G.R. Aase and D.R. Heiser. 2004. "Improving order-picking performance through the implementation of class-based storage." *International Journal of Physical Distribution & Logistics Management*, No.7, 534-544.
- Richards, G. 2014. *Warehouse Management 2nd Edition*. Kogan Page Limited, India.
- Rim, S.C. and Park, I.N. 2008. "Order picking plan to maximize the order fill rate." *Computers & Industrial Engineering*, No.55, 557-566.
- Tang, L.C. and E.P. Chew. 1997. "Order picking systems : batching and storage assignment strategies." *Proceedings of 1996 ICC & IC*, No. 3-4, 817-820.

AUTHOR BIOGRAPHIES



THANANY WASUSRI graduated with a PhD from the University of Nottingham, England in the field of manufacturing engineering and operations management. She is working at King Mongkut's University of Technology Thonburi. Her research of interest is logistics management, supply chain management and inventory management while scientific tools are simulation, optimization and multivariate statistics. Her e-mail address is : thananya.was@kmutt.ac.th



PRASIT THEERAWONGSATHON obtained a Master degree of Science in Logistics Management at King Mongkut's University of Technology Thonburi, Thailand. He is working with a footwear company in Thailand. His e-mail address is : tprasit@yahoo.com

Finance and Economics and Social Science

MODELING INFERENTIAL MINDS IN CONCEPTUAL SPACE

Carlos Barra
Independent Researcher
Las Pimpinelas 880, Concon, Chile
E-mail: c.barra@vtr.net

Enrique Canessa
Facultad Ingeniería y Ciencias, CINCO
Universidad Adolfo Ibáñez
Av. P. Hurtado 750, Viña del Mar, Chile
E-mail: ecanessa@uai.cl

Sergio E. Chaigneau
Escuela Psicología, CINCO
Universidad Adolfo Ibáñez
Av. D. Las Torres 2640, Santiago, Chile
E-mail: sergio.chaigneau@uai.cl

KEYWORDS

Conceptual Agreement Theory, evolution of concepts,
Agent-based modeling, abstract concepts.

ABSTRACT

We present an Agent Based Model (ABM) named MIMICS (Modeling Inferential Minds in Conceptual Space), which shows how a social group develops abstract concepts for achieving agreement in communication. Agents describe concepts by assigning properties to them based on learning and communication interactions, trying to develop a conceptual space that discriminates as much as possible between two concepts (i.e., they try to assign properties to concepts decreasing the overlap among the properties that describe them). Contrarily to concrete concepts, those properties come from the social group and not from objects' physical properties. The results show that agents in MIMICS develop abstract concepts that exhibit the same characteristics that are found in studies of real concepts: non-uniform frequency distributions of properties, inter-subjective variability and stable concepts that are useful for the simulated social group by providing agreement in communication.

INTRODUCTION

Concrete concepts are typically associated to physical properties. For example, a type of face could be characterized by properties such as a given nose length, eye separation, mouth width, etc. (Tversky 1977). The standard theory of concrete concepts holds that concepts are learned by observing category exemplars, extracting the relevant properties (Schyns et al. 1998), and estimating the frequency distribution of those relevant properties (Ashby and Alfonso-Reese 1995; Griffiths et al. 2011). This then allows organizing a semantic structure and making category judgments (e.g., How typical is a given exemplar of category X ? Does the exemplar belong to category X or category Y ? How central is property j for category X ?).

Perhaps one of the most important findings about concrete concepts, is that concepts relate only probabilistically to conceptual properties (Rosch 1973). This means that, among other things, two concepts that may be applied to a situation or object are not discriminable through a logical rule (i.e., by necessary and sufficient properties), but show a typicality structure instead (Rosch et al. 1976). Those exemplars that exhibit

frequent properties are more typical than those exhibiting less frequent properties (Rosch and Mervis 1975) (e.g., an ostrich would be a low typicality exemplar of the BIRD category, whereas a dove would be a typical exemplar). The typicality structure also means that an object can be a member of different categories, although to different degrees (e.g., a Chihuahua may be a low typicality exemplar of the DOG category, while simultaneously being a relatively more typical case of the PET category). Note here that the fact that an exemplar may belong to more than one category, implies that concepts must share properties to a certain extent (e.g., *being friendly to people* may be a property of the concept DOG, but also of the concept PET). Henceforth, we will refer to this as “conceptual overlap”.

In contrast to concrete concepts, relatively little is known about abstract concepts (e.g., freedom, democracy, personality). This is a problem, given that a large proportion of the concepts that we use are abstract concepts (estimated to be more frequent than concrete words, Rechia and Jones 2012). Though the standard concrete concept theory assumes that it is a valid description of all kinds of concepts, there is evidence that abstract concepts do not respond to the same characteristics.

When researchers study concrete or abstract concepts they frequently ask a sample of individuals to produce lists of conceptual properties (e.g., Wu and Barsalou 2009). However, for abstract concepts subjects do not produce physical properties. Rather, they produce verbal associations (e.g., for the concept EMERGENCY, we might obtain *danger* as one of its properties; Della Rosa et al. 2010). When these lists are coded and aggregated, non-uniform or non-homogeneous frequency distributions of conceptual properties are obtained (these are called norming studies).

Another difference between concrete and abstract concepts is the following. Though concrete concepts may be learned without supervision (e.g., Love 2002), it does not seem possible to learn an abstract concept without some kind of supervision. A concrete concept may be learned by perceiving a sequence of exemplars, while extracting common properties. It is dubious that the same could be achieved for an abstract concept. Though there is no empirical support for this claim, it is difficult to imagine a list of exemplars that would allow learning, e.g., the concept of SECURITY without some kind of feedback. Furthermore, many abstract concepts refer to internal states that are not directly perceptible (e.g.,

INTENTION, Wu and Barsalou 2009) and are therefore clearly different from concrete concepts.

A working hypothesis of the work we report here is that the two issues discussed above (i.e., supervision requirement and verbal associations) are related. If abstract concepts cannot be learned by observing exemplars, it is possible that learning them requires attending to the behavior of other members of the social group when they use the concept. For example, it is possible that some kind of explicit teaching of the concept is necessary for someone to learn which other concepts are associated with the focal concept. If so, the need for supervision may cause that the conceptual content of abstract concepts is basically verbal.

Another important feature of property frequency distributions (for abstract as well as concrete concepts) is that they show inter-subjective variability in conceptual content (Barsalou 1993). In fact, given that subjects produce different lists of conceptual properties, non-uniform distributions follow.

Given our discussion above, our general research question in the current work is why abstract concepts (just as concrete concepts) should produce a probabilistic structure of properties (i.e., verbal associates), given that abstract concepts seem not to be learned as concrete concepts are. In the next section, we develop a meta-theory that offers an explanation, which we later implement as an Agent Based Model (ABM). Note that although we don't strictly follow the ODD protocol (Grimm et al. 2006) to present the ABM, we comply with including most of the material suggested in it.

A META-THEORY ABOUT CONCEPTS

Our meta-theory implies several factors that operate simultaneously. First, we assume that people can interact with a concept in two different manners: either learning or using it. At any given moment, individuals should make a decision regarding how to interact with a concept (not necessarily a conscious one). We assume that this decision depends on how much an individual knows how to differentiate the concepts in question. Though there are several potential ways in which individuals could determine if they know a concept well enough to use it confidently (e.g., they could pay attention to feedback from others regarding whether they are using the concept correctly), in the current work we assume that individuals attempt to discriminate as much as possible the focal concept from other potentially applicable concepts. Thus, the lesser they are able to discriminate, the more they are prone to learn something new about the concept in question.

Given that it does not seem possible to learn an abstract concept merely by perceiving exemplars, it is likely that these concepts are learned through explicit information acquired from others. There are several ways in which this could happen (e.g., individuals could directly ask others about the associated properties and construct their own frequency distribution in a piecemeal fashion). In the current work, we assume that individuals can learn

the content of an abstract concept by asking explicitly if a property corresponds to a concept.

As a consequence of learning more about the focal concept, we assume that individuals increasingly tend to decide to use a concept rather than continue learning. Classically, it would be assumed that a concept coded in language would be used to make reference (e.g., the word "dog" could be used to refer to a specific dog or to the category DOG). In contrast, here we assume that when using abstract concepts, individuals are trying to understand the point of view of a conversational partner (i.e., if she conceptualizes a situation as a case of the focal concept or as a case of an alternative concept). Here, again, there are several ways in which this could happen (e.g., an individual could observe the conceptual content produced by someone and by an associative process could gain information about which concept is being used). In the current work, we assume that individuals first adopt a given point of view (i.e., they conceptualize the situation as a case of a given concept) and look for confirmation that their conversational partner has the same point of view (Chaigneau et al. 2012).

Though searching for confirmation is a strategy that will lead to errors (Nickerson 1998), in our work we assume that a social group could use it to keep useful concepts (i.e., those that allow inferring the likely mental state of others). Looking for confirmation is a very simple strategy, which is likely to be used more than sophisticated processes (e.g., disconfirmation), and that does not require assuming elaborate cognitive processing.

DESCRIPTION OF MIMICS ABM

We designed an ABM that implements a specific version of the meta-theory described above (MIMICS; Modeling Inferential Minds in Conceptual Space). This theory assumes specific solutions to the topics discussed above, though — as also discussed above — other solutions are possible. Just to refresh them, the topics are the distinction between learning and using a concept, how an abstract concept may be learned, and what does it mean to use an abstract concept. Thus, our specific goal is to test if the ABM formalization is able to produce the pattern of results exhibited by abstract concepts: probability distributions of properties, absence of an objective criterion to define concepts and inter-subjective variability, and, despite all that, stability and usefulness of concepts.

In MIMICS, agents play two types of roles: observers (*O*) and actors (*A*), and act as *O*s and *A*s depending on the type of interaction executed (see Table 1 and associated explanations). Regardless of the role, agents know there are two concepts that can apply to a situation (*C1*, *C2*), and that there are properties (*j*) that can describe them. They also have a finite universe of *P* potential properties $j \in \{0,1,2,\dots,P-1\}$ that can describe any of the two concepts $c \in \{1,2\}$. These are not properties in a traditional sense (i.e., they are not independently

discriminable perceptual features), but rather verbal tags associated to concepts.

Agents develop their concepts either communicating with other agents or learning from them. For each concept ($C1$, $C2$), agents keep track of the number of occasions (f_j^c) in which they have found property p_j when interacting with the given concept c , and of the number of times in which they have interacted successfully (d_j^c) with that property p_j relative to that given concept c (see below for an explanation of what constitutes a successful interaction). In general, the greater d_j^c is in relative terms, the greater the evidence is for that property p_j to belong to that concept c . Note that the potential property j becomes a known property p_j (we will explain this process later on).

MIMICS has 2 mechanisms for concept development based on social interactions:

One is an implicit mechanism in which O is not attempting to learn, but to decide if A is in the same mental state as he is (we call this process, communication). In this process, O believes it knows the concept sufficiently and that there is no need to continue learning it. Then, in the communication mode, O assumes that the situation can be described by $C1$ (or $C2$), and waits for evidence that A conceptualizes it similarly. Then, A selects a concept c and a property p_j that belongs to c (p_j^c), and offers that property to O . If that property is in O 's concept $C1$ (or $C2$), then O assumes that both agree about the situation's definition. Consequently, O increases d_j^c and f_j^c for that property in concept $C1$ (or $C2$), otherwise, O increases only f_j^c (not d_j^c) for that property in concept $C1$ (or $C2$). Note that agreement can be true (A is really also thinking of concept $C1$ (or $C2$)) or it can be illusory (A is not really thinking in $C1$ but in $C2$ (or not in $C2$ but in $C1$)). In other words, ABM agents cannot read other agents' minds, and can only infer their mind states based on the evidence.

The other mechanism is one of explicit learning. If O believes it needs to learn more about concept $C1$ (or $C2$), then it looks for more information. For that, O queries A with a c, j pair (i.e., asks whether j is a property of c in A 's mind). If the query receives a negative answer, then O increases f_j^c but does not increase d_j^c (i.e., signaling that j has been experienced, but that it is not part of the focal concept). If the query receives a positive answer, then O increases f_j^c and d_j^c .

For each property p_j^c (i.e., each j in each concept $C1, C2$), agent O computes a success probability ($SP_j^c = d_j^c / f_j^c$) for interacting with that property p_j^c in that concept c . SP_j^c is the probability, computed from an agent's own experience, that it can achieve agreement when using a given property p_j^c in a given concept c .

The information obtained in communication and learning is used by O for two things:

First, it uses it to decide to which concept to assign a property p_j^c . The probability of p_j^c being assigned, e.g., to concept $C1$, increases probabilistically as the normalized absolute difference between the SP s for property p_j^c also

increases (i.e., how much an individual knows how to differentiate the concepts). In general, as the number of successful interactions when using a property p_j^c increases (i.e., those interactions that produce agreement), the evidence for that property belonging to that concept, and not to an alternative concept, also increases. In other words, to assess the possibility of discriminating a property between both concepts, the agents use $|SP_j^1 - SP_j^2|$. A small absolute difference shows that property p_j^c is not very discriminable (i.e., it produces about the same success probability for both concepts). This value is normalized to obtain what we define as the discrimination probability:

$$DP = \frac{|SP_j^1 - SP_j^2|}{\forall p_j \max |SP_j^1 - SP_j^2|} \quad (1)$$

In eq. (1), the absolute difference in SP for property p_j^c for both concepts ($|SP_j^1 - SP_j^2|$) is divided by the maximum difference across all known properties in O 's mind (see below for an explanation of how an agent knows properties), so that DP will always fall in the $[0,1]$ interval. Thus, using DP , an O agent will probabilistically decide if it has enough information to discriminate. If that is the case, the discrimination process is accomplished by comparing SP_j^1 with SP_j^2 , so that if $SP_j^1 < SP_j^2$, p_j^c is assigned to $C2$ and it is withdrawn from $C1$ (and vice-versa). This built-in preference for clearly separable concepts has been posed as a basic tendency in human categorization. If possible, people prefer to form linearly separable categories (Blair and Homa 2001).

Second, O uses the information obtained in communication and learning to decide if its next interaction with an A should be in the learning or in the communication mode (as described earlier). To this end, O computes a measure of the "separation" that the properties p_j^c have achieved. In MIMICS, this measure is the average absolute difference of all the properties' SP . Based on this average, O probabilistically decides in which mode to interact. An increase in this average value, signals an increase in separation, and results in a decreased learning probability (LP) for O (i.e., the probability that O decides to continue learning). However, because an agent knows A properties for a given concept, LP is really computed as a representative average:

$$LP = \frac{1}{A} \cdot \sum_{\forall p_j} 1 - |SP_j^1 - SP_j^2| \quad (2)$$

Note that because agents discriminate and also decide to stop learning depending on their own experience with conceptual properties, inter-subjective variability follows naturally in MIMICS. Table 1 presents the pseudo-code of the learning and communication interactions. MIMICS randomly selects without replacement an agent from the list of all agents and that agent acts as O and O randomly selects another agent as an A , following the

actions defined in Table 1. This process is executed until all agents have been O_s , which constitutes a simulation step.

Table 1: Pseudo-code of Learning and Communication Interactions

OBSERVER O	ACTOR A
Preparation of the interaction	
1. Randomly selects an A actor from the rest of the agents 2. Randomly selects a concept c and property j from the P potential ones 3. Decides interaction mode: $\zeta Rdm(1) \leq LP?$	
$\zeta Rdm(1) \leq LP? = \text{FALSE (Learning mode)}$	
	4. Selects same c as O 5. If $c = \emptyset$ (<i>auto-learning</i>) randomly selects a property j from the P potential ones ($p_j^c = j$) and increments f_j^c and d_j^c . 6. Assigns $p_j^c = j$ of O 7. $\exists p_j^c \Rightarrow result = 1$ $\neg \exists p_j^c \Rightarrow result = 0$
8. Increments f_j^c 9. $result = 1 \Rightarrow d_j^c = d_j^c + 1$ 10. $SP_j^c = d_j^c / f_j^c$ 11. $d_j^c \geq 0 \Rightarrow p_j \in c$ 12. Discrimination Inference: $Rdm(1) \leq DP \Rightarrow$ a) $SP_j^1 < SP_j^2 \Rightarrow p_j \notin 1, p_j \in 2$ b) $SP_j^1 > SP_j^2 \Rightarrow p_j \in 1, p_j \notin 2$	
$\zeta Rdm(1) \leq LP? = \text{TRUE (Communication mode)}$	
	4. Randomly selects concept c and property p_j^c ($j = p_j^c$). 5. If $c = \emptyset$ (<i>auto-learning</i>) randomly selects a property j from the P potential ones ($p_j^c = j$) and increments f_j^c and d_j^c .
6. Assigns $p_j^c = j$ of A 7. $\exists p_j^c \Rightarrow result = 1$ $\neg \exists p_j^c \Rightarrow result = 0$ 8. Increments f_j^c 9. $result = 1 \Rightarrow d_j^c = d_j^c + 1$ 10. $SP_j^c = d_j^c / f_j^c$ 11. $d_j^c \geq 0 \Rightarrow p_j \in c$ 12. Discrimination Inference: $Rdm(1) \leq DP \Rightarrow$ a) $SP_j^1 < SP_j^2 \Rightarrow p_j \notin 1, p_j \in 2$ b) $SP_j^1 > SP_j^2 \Rightarrow p_j \in 1, p_j \notin 2$	

Note: LP see eq. (2), DP see eq. (1)

The initial conditions of a run consist in instantiating f_j^c , d_j^c and SP_j^c and the lists of properties that belong to concept $C1$ or $C2$ in each agent to null. Notably, interactions among agents depend on the interaction rules described above, and the only exogenous parameters are the number of agents (N) in a simulation and the number of potential properties for describing concepts $C1$ and $C2$ (P). This means that all results presented here can be

attributed to the interaction and decision rules (i.e., the meta-theory and MIMICS' solutions to the topics discussed earlier), and not to the way specific parameters were set in the experiments, except for N and P .

To simplify Table 1, we did not include the process by which agents know the properties. Agents know the existence of potential properties when properties are used in any of the interaction modes shown in Table 1. An agent knows property j by initializing $p_j^c = j$ and $d_j^c = f_j^c = SP_j^c = 0, \forall c$. There are two exceptional cases: 1) in communication mode, given that A presents to O a known property and does not receive anything from O , new properties are not incorporated by A as known and 2) when $c = \emptyset$ (i.e., when there is no conceptual content in c at initial conditions) for A in learning mode, A does an auto-learning process; that is, A not only initializes the property j , but also assigns it to c ($d_j^c = f_j^c = 1$) (i.e., property j becomes a known and assigned property p_j^c), which always happens at the beginning of a simulation run, when agents don't have any information or structure in their particular conceptual space. Finally, each run is ended when, at the social group system's level, the conceptual space structure is stable. That is determined when no further change is observed for the properties incorporated into concepts at the group's level. This occurs when the standard deviation of the average SP of both concepts (across all agents), calculated in a sliding window of 3,000 simulation steps, does not show significant variations; i.e., the standard deviation of the standard deviation of the average SP of both concepts is equal to or less than 0.004.

EXPERIMENTS AND RESULTS

Our general hypothesis is that a process based on social interactions (communication and learning), where conceptual properties come from the social group and not from objects' physical properties, is able to produce concepts characterized by non-uniform probability distributions and inter-subjective variability in conceptual content, while making minimal assumptions about agents' cognitive machinery. Specifically, we expect that, for a wide range of experimental conditions (N and P values), MIMICS will produce stable concepts that are useful for the simulated social group, but not at the expense of homogeneity in conceptual content (i.e., MIMICS should exhibit inter-subjective variability in conceptual content). Also, as a direct consequence of this, MIMICS should produce non-uniform frequency distributions of properties similar to those found in norming studies. For the experiments we set up $N = \{14, 40, 60\}$ and $P = \{10, 50, 100\}$. We selected those values for representing small, medium and large groups of agents and number of potential properties. Each of the nine experimental conditions was run 20 times and in all the graphs that show averages, these were computed using the output values of the 20 replications. We don't present std. deviations, given that they are very small and only would have cluttered the graphs. We performed an ANOVA for all the presented results (where suitable),

which indicates that all of them are highly statistically significant (all p -values ≤ 0.005). In the following paragraphs we present the results for concept $C1$, given that the ones for concept $C2$ are similar. For those interested in replicating our experiments, the program is available at <http://ccl.northwestern.edu/netlogo/models/community/>. You will have to search for the file MIMICS v-CSI.netlogo, found under the March 2016 heading, and download it to your computer. Then you need to download and install the Netlogo platform, version 4.0.4 at <http://ccl.northwestern.edu/netlogo/oldversions.shtml>. To assess the usefulness of concepts, we use the probability of true ($p(a1)$) and illusory agreement ($p(a2)$) per Conceptual Agreement Theory (CAT, Chaigneau et al. 2012). According to CAT, when human beings talk about abstract concepts (e.g., democracy, political views, masculinity, personality traits), they try to infer agreement, i.e., to infer whether other people’s mind-content is similar to their own content or not. To illustrate, imagine two individuals, O and A , that are having a conversation about a given topic, and that O has a hypothesis $C1$ about how entity x is being jointly conceptualized (i.e., that they are talking about x as an instance of $C1$). However, because concepts are events in individual minds, O can only infer whether $C1$ is the case for A or not. To make this inference, O observes A , and when A describes x as having a property p_i , O evaluates if p_i is consistent with $C1$ in her mind or not. If it is consistent, then O infers that A is also talking about x conceptualized as $C1$. If A is in fact talking about x conceptualized as $C1$, then this is true agreement (event $a1$ and its probability is $p(a1)$). If A is talking about x conceptualized as $C2$, then illusory agreement happens (event $a2$ and its probability is $p(a2)$). Note that this situation corresponds to the idealized communication interaction shown in Table 1. In MIMICS, to compute $p(a1)$, each time agents engage in a communication interaction and both are using concept $C1$, a counter f_{a1} is incremented. On the other hand, if O is using concept $C1$ and A is using $C2$, a counter f_{a2} is incremented. If agents infer agreement and that is actually true agreement (both agents are actually thinking of $C1$), then a counter $a1$ is incremented. Contrarily, if agent O is thinking of $C1$ and agent A is thinking of $C2$, then a counter $a2$ is incremented. Calculating $p(a1)$ and $p(a2)$ amounts to dividing $a1$ by f_{a1} and $a2$ by f_{a2} . Given that concepts should afford a $p(a1)$ larger than $p(a2)$ to be useful in communication among members of a group (i.e., more true than illusory agreement; Chaigneau et al. 2012), we should observe the same in MIMICS’ outputs. As Figure 1 shows, that is the case. For all the nine experimental conditions, always $p(a1)$ is larger than $p(a2)$, which means that agents develop a conceptual space that promotes true agreement in communication.

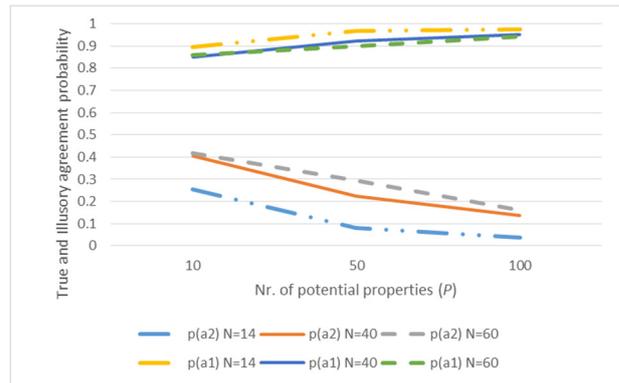


Figure 1: Avg. $p(a1)$ and $p(a2)$ for the 9 Experimental Conditions

On the other hand, although a high true agreement is reached, agents exhibit inter-subjective variability in conceptual content. To illustrate this, we can inspect the properties assigned to concept $C1$ by two agents in a given experimental condition ($N = 40$, $P = 100$). For example, agent 0’s content for $C1$ is [2 4 5 8 10 16 17 43 63 67 68 71 74 77 78 90 97], whereas agent 10’s content is [8 10 27 33 34 38 62 64 68 75 77 85 97]. To generalize this claim, Figure 2 shows the frequency distribution of the properties across agents for $C1$, for the same experimental condition. It can be seen that the distribution is non-uniform (which also supports our assertion that MIMICS would produce non-uniform frequency distributions of properties). Given that the distribution is non-uniform, the only way that may happen is if agents have diverse conceptual contents. To more generally back up our claim, Figure 3 shows MIMICS’ outputs k_i and s_i . Variable k_i corresponds to the total number of properties for concept $C1$ in a population of individuals, and s_i to the average number of properties coherent with concept $C1$ in an individual’s mind (Chaigneau et al. 2012).

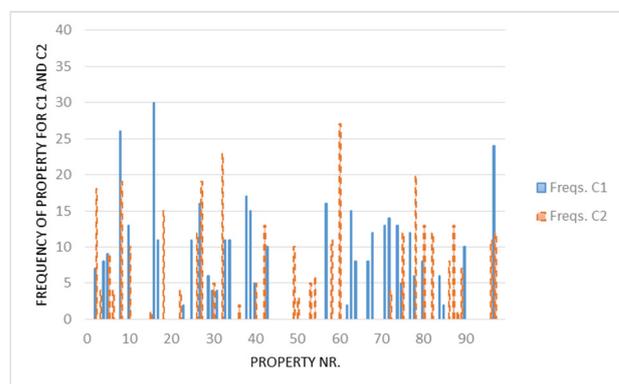


Figure 2: Frequency Distribution of Properties across Agents for Concept $C1$ ($N = 40$, $P = 100$)

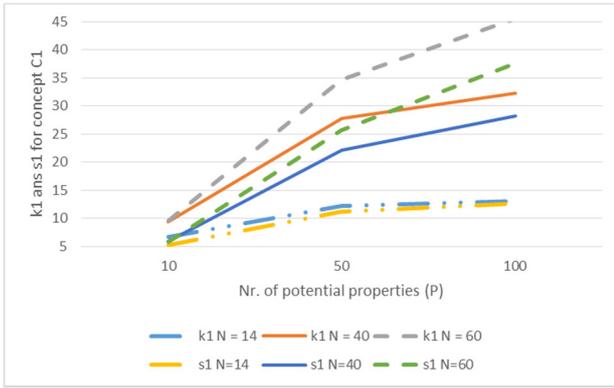


Figure 3: Avg. k_I and s_I for the 9 Experimental Conditions

From Figure 3, we can see that for all the experimental conditions, k_I is always larger than s_I . That may happen only if the number of properties assigned to $C1$ in agents' minds is smaller than the total number of properties assigned to $C1$ across all agents, which proves that inter-subjective variability in conceptual content must exist. As already discussed, Figure 2 supports our claim that MIMICS would produce non-uniform frequency distributions of properties similar to those found in norming studies. To generalize this finding to all the nine experimental conditions, Figure 4 presents the standard deviation for the properties for $C1$ (i.e., for the numbers that represent those properties).

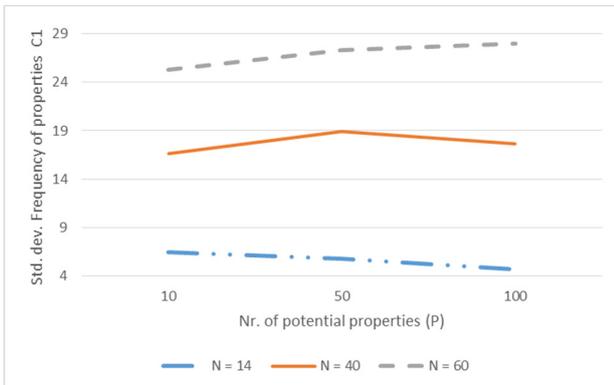


Figure 4: Avg. Std. Deviation of Properties of $C1$ for the 9 Experimental Conditions

The standard deviations different from zero confirm that for all experimental conditions, the frequency distributions of properties are non-uniform.

Finally, as pointed out earlier, another characteristic of these frequency distributions obtained in norming studies, is that the properties, which describe concepts, exhibit some conceptual overlap (i.e., some properties describe more than one concept, as one can see from Figure 2). To generalize that finding, Figure 5 shows a normalized RMSE of the frequencies of the properties that describe $C1$ and $C2$ in MIMICS. The normalized RMSE was calculated by first dividing the frequency of each property that describes $C1$ and $C2$ by the respective maximum frequency. Then, the sum of the squared difference between the normalized frequencies of $C1$ and

$C2$ was divided by the number of frequencies that are larger than zero in both concepts, and the square root of that mean is the RMSE.

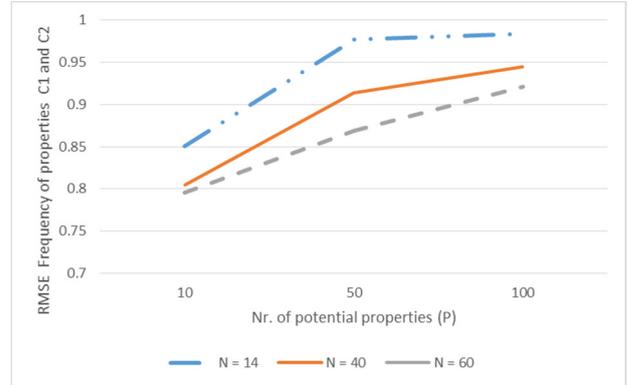


Figure 5: Avg. Normalized RMSE of Properties of $C1$ and $C2$ for the 9 Experimental Conditions

This form of calculating the RMSE assures obtaining a superposition index between the $C1$'s and $C2$'s property frequency distributions that is comparable across different N and P . That will be an important issue when further analyzing MIMICS' results in future work. Note from Figure 5 (and also from Figure 2), that under all nine experimental conditions, a RMSE above zero, implies that indeed there exists conceptual overlap between concepts, just as has been found in empirical studies.

DISCUSSION/CONCLUSIONS

As we discuss in the introduction, when subjects are asked to produce conceptual content for abstract or concrete concepts, non-uniform frequency distributions of properties are obtained. Also, there is inter-subjective variability in conceptual content. For concrete concepts, these properties are descriptors of the concrete objects that belong to the category, and non-uniform distributions occur because some properties are more frequent than others in the exemplars that belong to the category (e.g., most dogs bark). In contrast, for abstract concepts properties are verbal or conceptual associations, and it is unclear why non-uniform property distributions and inter-subjective variability should be obtained.

In the current work, we present MIMICS, which is a theory of how a social group develops a system of abstract concepts. MIMICS makes three important assumptions about abstract concepts (widely supported by the literature that we have cited throughout this paper). First, it assumes that abstract concepts are states of mind or points of view about a situation. As such, they cannot be directly observed and need to be inferred. Second, it assumes that individuals are motivated to know if other individuals share their own particular point of view. Third, it assumes that—as is true of concrete concepts—people attempt to learn linearly separable concepts. Noteworthy about MIMICS, is that conceptual content develops from social interaction and not from the environment's structure. There are two kinds of interactions: learning from other group members, and

communicating (i.e., using conceptual properties to infer a conversational partner's state of mind).

In our computational experiments with MIMICS, we found that for a wide range of experimental conditions (i.e., combinations of N and P values), MIMICS reproduces the type of results that are obtained in conceptual norming studies. MIMICS' rules of interaction are successful in producing non-uniform property frequency distributions, concepts that are not neatly discriminated, and agents with non-homogeneous conceptual content, though agents do not extract this structure from an environment. Just as importantly, MIMICS produces concepts that, despite inter-subjective variability, allow communication. Concepts developed by MIMICS allow agents more often than not to correctly infer conceptual agreement with other agents.

There are several situations in which a researcher may need to explore the use of abstract concepts in a social group. An anthropologist may want to know whether a social group holds a shared view on a socially relevant topic (e.g., how political parties are characterized); a marketing expert may want to know whether a social group holds a shared view on a brand's image). We envision using MIMICS and the theoretical insights derived from our simulations as providing tools to analyze problems such as these.

There are many other conclusions that can be drawn from our results. However, they are beyond the scope of this paper and will be part of our future work with MIMICS. What we want to stress here is that MIMICS shows that abstract concepts may be advantageously viewed as devices developed by a social group to allow agreement and mind-reading.

ACKNOWLEDGMENT

This work was supported by FONDECYT (Fondo Nacional de Ciencia y Tecnología of the Chilean Government) grant Nr. 1150074 to the last two authors.

REFERENCES

- Ashby, F.G. and Alfonso-Reese L.A. 1995. "Categorization as Probability Density Estimation." *Journal of Mathematical Psychology* 39, No 2, 216-233.
- Barsalou, L. W. 1993. Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. C. Collins, S. E. Gathercole, & M. A. Conway (Eds.), *Theories of memories*. London: Erlbaum. pp. 29-101.
- Blair, M. and Homa, D. 2001. "Expanding the search for a linear separability constraint on category learning." *Memory and Cognition* 29, No8, 1153-1164.
- Chaigneau, S.E.; Canessa, E.; and Gaete, J. 2012. Conceptual agreement theory. *New Ideas in Psychology* 30, Nr 2, 179-189.
- Della Rosa, P.A.; Catricalà, E.; Vigliocco, G.; and Cappa, S. 2010. Beyond the abstract-concrete dichotomy: mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods* 42, 1042-1048.
- Griffiths, T.L.; Sanborn, A.N.; Canini, K.R.; Navarro, D.J.; and Tenenbaum, J.B. 2011. "Nonparametric Bayesian models of categorization." In E.M. Pothos and A.J. Wills (Eds.) *Formal Approaches in Categorization* 2011, pp. 173-198. Cambridge: Cambridge University Press.
- Grimm, V. et al. 2006. "A standard protocol for describing individual-based and agent-based models". *Ecological Modelling* 198, 115-126.
- Love, B. C. 2002. "Comparing supervised and unsupervised category learning." *Psychonomic Bulletin & Review* 9, 829-835.
- Nickerson, R.S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology* 2, Nr 2, 175-220.
- Recchia, G. and Jones, M.N. 2012. "The semantic richness of abstract concepts." *Frontiers in Human Neuroscience* 6, Art. 315, 1-16.
- Rosch, E. 1973. "On the internal structure of perceptual and semantic categories." In T. E. Moore (Ed.), *Cognitive Development and the acquisition of Language* 1973. New York: Academic Press.
- Rosch, E. and Mervis, C.B. 1975. "Family resemblances: Studies in the internal structure of categories." *Cognitive Psychology* 7, 573-605.
- Rosch, E.; Simpson, C.; and Miller, R.S. 1976. "Structural bases of typicality effects." *Journal of Experimental Psychology: Human Perception and Performance* 2, Nr 4, 491-502.
- Schyns, P.G.; Goldstone, R.L.; and Thibault, J.P. 1998. "The development of features in object concepts." *Behavioral and Brain Sciences* 21, Nr 1, 40-41.
- Tversky, A. 1977. "Features of similarity." *Psychological Review* 84, No 4, 327-352.
- Wu, L.L. and Barsalou, L.W. 2009. "Perceptual simulation in conceptual combination: Evidence from property generation." *Acta Psychologica* 132, 173-189.

AUTHOR BIOGRAPHIES

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.

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.

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 inter-subjective agreement.

Individual-level Simulation Model for cost benefit analysis in healthcare

Nagesh Shukla^{1*}, Vu Lam Cao¹, Van Hoang Phuong², Marian Shanahan², Allison Ritter², Pascal Perez¹

¹SMART Infrastructure Facility
Faculty of Engineering and Information Sciences
University of Wollongong
NSW, Australia 2522

* Email: nshukla@uow.edu.au

²Drug Policy Modelling Program
National Drug and Alcohol Research Centre
University of New South Wales
NSW, Australia 2052

KEYWORDS

Individual-level simulation model; Health Economics; Illicit Drug Use.

ABSTRACT

Illicit drug use creates significant burden at societal, family and personal levels. Every year substantial resources are allocated for treatment and the consequences of illicit drug use in Australia and around the world. Heroin is one of the major forms of illicit drugs. Several independent heroin treatment strategies or interventions exist and state-of-the-art research demonstrates their efficacy and relative cost-effectiveness. However, assessing total potential gains and burden from providing all treatment interventions or varying the mix of heroin treatments has never been attempted. This paper proposed an individual-level simulation model (ISM) which addresses net social benefit over a lifetime that can accommodate the complexity of individuals going in and out of multiple treatments and their corresponding costs and benefits arising from different treatments during the life-course of heroin users in the context of New South Wales (NSW) Australia. This model is intended to serve as an effective tool for economic evaluation and policy making in the illicit drug area in Australia. The validity of the model has been assessed by comparing short term outcomes or examining the status of participants at a various points of time predicted from the model with other data sets that were not used to parameterise the model. Initial model results have been also presented to highlight different types of scenario analysis that can be conducted in future.

INTRODUCTION

This paper deals with the development of novel individual-level simulation model (ISM) for health care decision making. The model is developed for modelling range of treatments available for illicit drug users. The model is simulated for long term (generally lifetime of drug users) to evaluate treatment services. Following paragraphs deal with the area of illicit drug use. Governments, non-governmental organizations (NGOs) and International Organizations worldwide invest hundreds of billions of dollars in health care projects. Australia spends around 10% of its GDP or AUD 100 billion per year in recent years in health care WDI (2012). In the area of illicit

drug spending, Australian federal and state governments spend about AUD 1.7 billion per annum in prevention, treatment, harm reduction and law enforcement to combat illicit drugs. There is an increasing pressure from both the government and the public to know whether the current spending is optimal or what needs to change to increase the benefits of spending. This is particularly important for complicated policies where there are many external costs and benefits, and as such; there are diverse views about the value of the interventions.

Existing research demonstrates efficacy and relative cost-effectiveness for individual heroin treatments, such as pharmacotherapy maintenance. "Cost of illness" studies have estimated the total social burden related to all illicit drugs, and have been important in communicating this burden. But these studies do not provide evidence on the total potential gains from all interventions. And neither of these approaches can be used to value the net benefit, over the lifespan, of providing a system of heroin treatment interventions. There is a pressing need to demonstrate whether the existing combinations of heroin treatment interventions are a good investment for government. The aim of this paper is to develop a modelling approach which can assess the net social benefit of current heroin treatment strategies, and compare different combinations of treatment alternatives through modelled scenarios. This will lead to better informed policy decisions about the mix and type of treatments. The critical methodological issue is the choice of modelling approach. The model needs to capture recurring events over time as well as reflect alternative trajectories for individuals who use heroin. The chosen model is a micro-simulation model, also referred to as an ISM. It depicts events and outcomes at the level of the individual. The ISM enables 'memory' for each individual of such things as the length of heroin use, past treatments and incarcerations. This paper describes the rationale for an ISM and provides the detailed methodology employed to develop the ISM of heroin careers.

LITERATURE REVIEW

The decision problem of heroin use with recurring events such as abstinence, crime, incarceration, and treatment over time can be modelled using state transition models (STM). STMs consist of set of mutually exclusive and collectively exhaustive health states. Individuals can transition among

these health states based on prescribed transition probabilities. Interactions between individuals are ignored in STMs. One of the predominantly used STM in substance abuse literature is cohort-based STM also known as Markov model (Schackman et al. (2011)). Cohort-based STMs are relatively simpler to develop when the number of health states is not large. These models are restricted by the Markovian assumption, where transition probabilities do not depend on individual history or memory (i.e. past health states or state duration) Siebert et al. (2012). In case of heroin use, it is recognised that individuals history of incarceration, treatment, and length of time in treatment has an effect on state transition of an individual (Hser et al. (2004), Zhang et al. (2003)). Therefore, it is necessary to include the information about individual history while modelling heroin users based on STM. Cohort-based model can handle memory by having additional health states to include history, however, it often results in a very large model which is difficult to handle. On the other hand, individual-level simulation model (ISM) or individual-level STMs are not restricted by the Markovian assumptions as they simulate individuals history by using tracker variables. This greatly reduces the number of health states required.

ISMs better represent heterogeneity among individuals in complex modelling scenarios such as illicit drug use. ISMs can model individual characteristics as continuous variables whereby future decisions depend on current and past individual history. In case of cohort-based STMs, individual characteristics needs to be categorized to make separate health states Siebert et al. (2012). The ISMs can easily handle individual specific time steps as individuals in the model are simulated one at a time.

MODEL OVERVIEW

The ISM model in this paper starts with the population of individuals who have ever used heroin (previously and currently) for the New South Wales (NSW) state of Australia. These individuals are distributed in various health/treatment states (eg, abstinence, irregular use, dependent use, various treatment, prison and death states). Overall model working is conceptually represented in Fig. 1. There are six model components which are conceptually defined, namely, the initial population, health states, state transitions, costs and outcomes (these will be attached where relevant to being in a given state i.e. treatment, prison, societal costs of crime), and net social benefit.

The model starts with the initial population of current heroin users and heroin abstainers. This population of individuals are transitioned from one health state to other using predefined (individual based) state transition probabilities. After each state transition, outcomes such as heroin use, crime committed; and resource implications are computed. This process is repeated at each time step (where time step is defined as the length of stay in each state, individually driven) until the end of simulation time period is reached. Each year, a sub-population of new drug initiators is added to the current population to include new drug users, along with a mortality rate which exists for the model). Finally, net social benefit is computed based on the outcomes of the simulation model. Following section provides more detail on each of the model components.

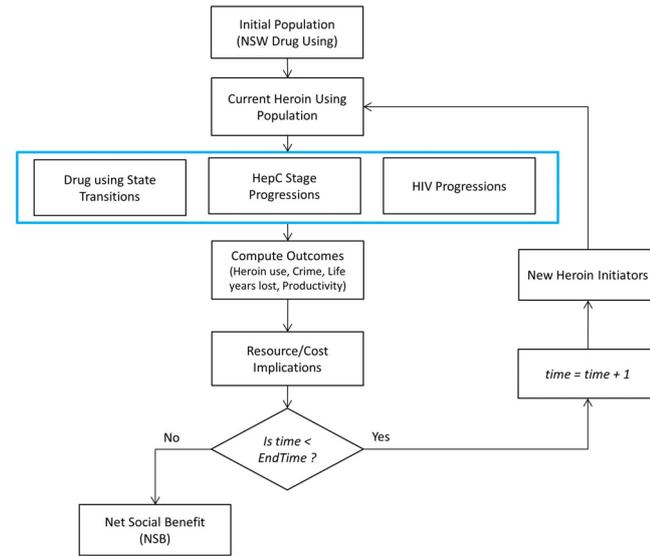


Fig. 1. Model Overview

Initial Population

The initial population in the current model was estimated based on the current NSW heroin using population. This includes those currently abstinent, those in treatment subgroups as well as those currently in heroin using subgroup. The characteristics of the initial population that were selected were age, gender, treatment history (number of episodes and duration of treatment), HIV and Hepatitis-C status; which were obtained from a number of data sources.

States

In the ISM, we have used three important locations (stages) in the drug using individuals trajectory : i) in community, ii) in prison, and iii) death. The first two stages are considered in this study to model the cost, benefit and treatment variations in drug using population. Exit from the model occurs if alive at age 60, death from drug related or non-drug related causes. Hence the total number of states is provided in Table I.

TABLE I. TOTAL NUMBER OF HEALTH STATES BASED ON HEROIN USE AND STAGES

State	Stage	Description
Abstinence(S1)	COMMUNITY	not using
Irregular use(S2)	COMMUNITY	use irregularly
No Treatment & Use(S3)	COMMUNITY	dependent users with no treatment
Withdrawal(S4)	COMMUNITY	Withdrawal treatment
Residential rehab.(S5)	COMMUNITY	rehabilitation
Pharmacotherapy (S6)	COMMUNITY	provision of safe opioid
Counselling Only(S7)	COMMUNITY	psychological therapy
Prison No Treatment(S8)	PRISON	not in treatment
Treatment in Prison(S9)	PRISON	in treatment
Death or 60+ years old	DEATH	death from all causes

Transition Time

In this model, an approach which provides heterogeneous time to transition for each individual is used. We have used length of stay (LOS) distributions for each individual. This is due to the fact that previous research has shown that LOS is highly predictive of subsequent drug use outcomes (Zhang et al. (2003), Hser et al. (2004)).

The distributions of LOS of an episode is created and assigned to each individual in the model. The LOS varies depending on individual characteristics (such as previous treatment episodes, age, gender, amount of drug use, state). The LOS based on individual characteristics were used only where sustained evidence exist (i.e. at least two empirical evidences to confirm). See Table II for data sources. Two type of survey dataset such as Alcohol and Other Drug Treatment Services National Minimum Data Set (AODTS-NMDS) and Australian Treatment Outcome Study (ATOS) have been used.

TABLE II. SUMMARY OF DATA SOURCES FOR LOS ESTIMATION

State	LOS data sources
S1	ATOS Dataset; Simpson and Marsh, 1986; Termorshuizen et al. (2005); Hser et al. (2007); Shah et al. (2006); Nosyk et al. (2013)
S2	ATOS Dataset; Gronbladh & Gunne (1989); Coffin & Sullivan (2013)
S3	ATOS Dataset; Bell et al. (1997)
S4	ATOS Dataset; (AODTS-NMDS) Dataset
S5	ATOS Dataset; AODTS-NMDS Dataset
S6	ATOS; Burns et al. (2009)
S7	ATOS Dataset; AODTS-NMDS Dataset
S8, S9	average sentence for the crime type.

State Transitions

Once individuals in each state finish their assigned LOS in a state, they transition to other state based on transition probability functions. These functions are dependent upon the individuals attributes. In the model, these probabilities were estimated based on number of data sources. Australian treatment and outcome study (ATOS) dataset, MIX dataset, and relevant literature were used to estimate these probability functions.

Costs and Outcomes

The individuals in the model transition from one state to another and in this process costs and outcomes (also referred to as rewards) are accrued. The resources were attached to individual when they were in certain states and are referred to as state awards. For example, while in S4 (withdrawal in the community), the cost per day of withdrawal is attached; similarly cost was attached to residential rehabilitation (by days in RR); pharmacotherapy (by days in OTP); counselling (by days); prison (days); treatment in prison (days in OTP). Average unit costs by treatment type were applied.

The variation in resource use is driven by the length of time in a state. There are resources attached to some transitions (transition awards) i.e. transitioning into prison would incur the costs of the police and court. Total costs include the following components: (i) value of life-years (saved, or lost), (ii) treatment costs; (iii) other health care utilisation (i.e. treatment for specific diseases such as HIV, Hepatitis C), (iv)

crime and criminal justice system costs. Total benefits include: earnings due to individuals returning to work after successful treatments.

Net social benefit

Once the costs and benefits have been calculated, the criterion for assessing the overall efficiency of an intervention is the Net Social Benefit (NSB) Feldstein (1964). The main costs and benefits will be calculated based on participants life time trajectories, during which they may be employed, use health care services, contract HIV or Hep C, commit crimes and go to prison. The base model will characterize the status-quo of behaviours of heroin users and calculate the base net social cost-benefit of current heroin treatment arrangement in NSW; different policy options can be explored to find the best combination of treatments that gives the highest net social cost-benefit.

DATA SOURCES FOR THE MODEL

Initial Population

The initial population for the model provides an estimate for the number of individuals within each of the 9 states in the model. States S10 was not included in the initial population as this state relates to death. The figures for each state were derived from various recent data sources discussed in this section (where possible from years 2012-2013). The final numbers have been rounded, to reflect the lack of precision in the estimates (see Table III).

TABLE III. FINAL STARTING POPULATION NUMBERS

State	Starting number
S1	22,000
S2	53,000
S3	16,500
S4	20
S5	80
S6	17,500
S7	100
S8	1,100
S9	1,100
TOTAL	111,400

Transition Functions

The proposed model simulated life trajectories of heroin users in NSW based on the transition matrix, which lays out all possible transitions from one state to the other state over participants life time. The possible transitions are determined by the availability of treatment modalities in NSW and characteristics of heroin use such as initiation, developing to dependent use, participating in treatment, abstinence, relapsing and so on. A participant makes transition to one of mutually exclusive states once he/she has completed a LOS for the episode. Multiple datasets and published literature were used or combined to estimate the parameters for transition functions.

Transitions from community states to community states: ATOS was the primary dataset to estimate transition probability functions from community states to community states. In addition, evidence from literature was used together with ATOS estimates to determine the final transition probabilities.

Transitions from community states to prison states: Participants in S1, S2, S3, S4, S5, S6 and S7 are allowed to make transitions to prison states because of their previously committed crimes. The primary data sources to estimate the prison transition probabilities were the published results from a study about engagement with criminal justice system among opioid dependent people in NSW (Degenhardt et al. (2013)).

The alleged individuals (based on crime committed rates) can be proved or pledged guilty. The weighted average prison terms in the local and district courts by categories of offences were used to assign length of stay to individuals who enter prison (BOCSAR Court Data Report, 2012). All transition probabilities of imprisonment from S1-S7 are dependent on age, gender, and state history.

Transitions from prison states to community states: Individuals can move to one of states in community on the completion of their stay in prison.

Mortality

The model considered variety of mortality rates based on the individual characteristics such as length of stay in the current state and history of previous state. The data for these mortality rates were derived from PHDAS, ATOS, and RCT sample. The specific rates (crude mortality rates (CMR) per 1000 person years (PY)) used in the model are illustrated in Table IV.

TABLE IV. MORTALITY RATES FOR EACH STATE

State: In	State: From	Time	CMR (/1000PY)	
S1	Any	fixed	5.3 (5.0 to 5.6)	
S2	Any	fixed	5.3 (5.0 to 5.6)	
S3	S1, S4, S5, S7	1st week	17.4 (11.7-25.0)	
		2nd week	20.1(13.8-28.4)	
S4	S8	1st 2 weeks	59.5(41.3-83.6)	
		out of prison		
		Other time in S3	11.5(11.1-12.0)	
S5	S2, S6, S10	fixed	6.0(5.76.4)	
S6	anywhere	fixed	6.0(5.76.4)	
S7	S3, S4, S5, S7, S8	1st week	39.5(31.9-48.8)	
		S10	1st week	10.9 (4.0-23.8)
		S6	2nd week	17.0 (11.8-23.6)
		S6	Other time in S6	5.6(5.2-5.9)
S8	S3,S4, S5, S6	fixed	6.0(5.7-6.4)	
S9	anywhere	fixed	2.7(2.0 - 3.7)	
S9	anywhere	fixed	0.7(0.3 to 1.2)	

Estimating costs

Treatment costs: The primary source for resource use information for treatment was extracted from individual care plans developed for the National Drug and Alcohol-Clinical Care and Prevention (DA-CCP). Care plans for opioid substitution, withdrawal, residential rehabilitation, and counselling for populations aged 18 to 64 were used. Information on staff type and time, pharmaceuticals, diagnostics, overhead and administrative allocations were obtained from respective care packages for each treatment. Once identified, resources were costed in 2012 AUD, sourced from NSW Wages and Salaries, Medical Benefit Schedule, Pharmaceutical Benefits Schedule (NSW Health 2015, PBS 2015, MBS (2015)). Costs were estimated for an episode, and then a cost per day was calculated for use in the model.

Criminal justice system costs: Unit costs were obtained from multiple sources and were adjusted to 2012 AUD as necessary. The average cost of a day in prison in NSW was sourced from the Report on Government Services as was the average cost per charge in the Magistrates Court (Productivity Commission 2015). Social costs of crime were obtained from an Australian report which included the intangible losses, property losses, and medical costs by type of offence (Russell et al. (2013)). The costs of policing were also estimated (Byrnes et al 2012), with annual non-capital expenditure on policing (Productivity Commission 2015).

Value of Statistical life year: According to Access Economics (2008) the mean of value of a life from these 17 studies was \$5.7 million (range \$0.9 to \$28.4 million 2006 AUD). After further analysis the recommendations from this report suggest using \$6.0 million 2006 AUD (range \$3.7 to \$8.1 million). After adjusting to 2012 AUD with the CPI, the value of a life was estimated to be \$7.0 million (range\$5.87-\$8.34 million). This value was annuitized over 80 years with a 3% discount rate. Then the total value over the remaining expected lifespan was calculated.

Other health care utilization costs: Estimates of other health care utilisation were estimated as daily costs. These included utilisation of inpatient, emergency department, outpatient services, general practitioners, specialists, and ambulances by model state were obtained (NHC 2015, DoH 2014, ASNSW 2015). These costs were then applied in the model as relevant.

Benefits

Individuals who are in states S1, S2, S3, S6 and S7 can be employed. The probability of employment increased if individuals maintained longer duration in abstinence and pharmacotherapy treatment. It was assumed that the longer duration of abstinence and in maintenance treatment increased participants probability of employment. Probabilities of employment were derived from ATOS, MIX, and NSW Labour Statistics 2012. If the participant is employed, the benefit was calculated as equal to days of employment times earnings per day; if the participant is unemployed, the benefit was equal to zero. The mean weekly earnings by gender and age from the Employee Earnings Statistics published by the Australian Bureau of Statistics was used to calculate the total earnings in a state.

HIV and Hep-C prevalence among heroin users

The Hep-C prevalence and incidence rates were estimated from the literature (Shand et al. 2014) and the resulting calculated indicated that there were 62,056 individuals that were Hep-C infected. Then this number were split by applying age and gender distribution. Further, different stages of Hep-C infection was also considered. Another estimation indicated that there will be 670 to 840 new cases of Hep-C each year.

It is estimated that there are between 12,500 and 15,000 cases of HIV in NSW (NSW HIV Strategy 2012-2015). The number of cases attributable to IDU is approximately 2%; this is based on data that the prevalence of HIV is 1-2% among people attending NSPs (HIV Annual Surveillance Report, 2014). This equates to 300 existing cases of HIV in NSW

that are the result of drug use. Further, 458 new cases of HIV in NSW in 2012, 18 of these cases were attributable to drug use (16 male; 2 female).

MODEL VALIDATION AND RESULTS

The proposed simulation model was written in Java using the Eclipse IDE. The computer program was written following modular approach (object orient programming) for model components. Several verification tests were defined for each of the modules to verify the intermediate simulation outputs. The logic for transition of individuals within the conceptual model was matched with the output from the simulation model. Additionally, we have randomly selected some individuals in the simulation model and traced their behavior and matched that against our conceptual model logic. We have verified individual transitions within the model. The transition summary from the model was compared with the expected transition summary which was derived after manually inputting the values into the transition functions.

In terms of validation of the simulation model, we have used face validation and cross-validation. In face validation, we have conducted review meetings with the advisory group consisting of experts from the illicit drug field. The individuals, discrete health states, transition probabilities, costs, mortality, infection rates, and other parameters were also developed in conjunction with domain experts along with making simplifying assumptions to create the conceptual model. We also performed cross-validation by running the baseline simulation model and comparing its output with the published literature.

Cross-validation to existing datasets and published literature: The datasets and published literature which were not used for model building were used for cross validation exercise. Data on the overall behavior of illicit drug users were used for model validation. Following are some of the model validation results:

- 1) We have compared the simulation results to expected total number of individuals in S6. Estimates are used from NOPSAD data for the OTP numbers in treatment (on census day). According to this dataset, there were 18,715 individuals in OST treatment in NSW on the census day in 2012. It is expected that this number will grow at about 3% overtime assuming it will follow the previous trend. We compared this number with the number of individuals in state S6 in various simulated years. The rate of increase in the heroin users in S6 (OST) from simulation model was 2.929% (1.03 - 3.69 at 95% CI).
- 2) Larney & Indig (2012) estimated that the proportion of opioid-dependent prisoners receiving OST treatment was 43%. Based on the model the percentage of opioid-dependent prisoners receiving OST treatment in prison was 41.07 (95% CI 40.48 – 41.65).
- 3) According to the published literature, 60% of community participants are in some type of treatment. From the simulation model, the proportion of dependent users in some type of treatment was 48.71% (47.42-55.89 at 95% CI).
- 4) Mortality rate for opiate related deaths in NSW, according to Roxburgh and Burns (2012), was compared with the rate of drug related deaths in the

proposed model. The drug related deaths in NSW stated by Roxburgh and Burns (2012) for 15 - 54 years old population was 198 (at 95% CI 112 - 283) from 1999–2008. This means that the drug related mortality rate was 0.0029 (=283/97,592) - 0.00114 (=112/97,592) (at 95% CI). From the simulation model, the rate of drug related deaths (for 18 to 54 years old) was 0.00308 (0.002730 - 0.003446 at 95% CI), which is consistent with the rates reported by Roxburgh and Burns (2012).

- 5) Abstinence rate for 11 year follow up for dependent heroin users from ATOS study (Teesson et al. presentation) was used for validation. It is stated in the study that percentage of heroin dependent population decreases from 97.6 % in baseline year to 15.1 % in 11th year (85% decrease approx.). Based on the simulation model, we observed 47% decrease in heroin dependent population after 11 years. Similarly, percentage of dependent population in current treatment has decreased by 74% (after 11 years of simulation) compared to 46% stated in the Teesson et al. presentation.
- 6) Mortality rate for opiate related deaths in NSW, according to Roxburgh and Burns (2012), was compared with the rate of drug related deaths in the proposed model. The drug related deaths in NSW stated by Roxburgh and Burns (2012) for 15 - 54 years old population was 198 (at 95% CI 112–283) from 1999–2008. This means that the drug related mortality rate was 0.0029 (=283/97,592) - 0.00114 (=112/97,592) (at 95% CI). From the simulation model, the rate of drug related deaths (for 18 to 54 years old) was 0.00308 (0.002730 - 0.003446 at 95% CI), which is consistent with the rates reported by Roxburgh and Burns (2012).
- 7) Abstinence rate for 11 year follow up for dependent heroin users from ATOS study (Teesson et al. presentation) was used for validation. It is stated in the study that percentage of heroin dependent population decreases from 97.6 % in baseline year to 15.1 % in 11th year (85% decrease approx.). Based on the simulation model, we observed 47% decrease in heroin dependent population after 11 years. Similarly, percentage of dependent population in current treatment has decreased by 74% (after 11 years of simulation) compared to 46% stated in the Teesson et al. presentation.

Above-mentioned results indicate that the model results are consistent with the published studies. After validation, the results of the model which has been run for 25 years is obtained. The initial population was evolved in time to simulate population aging, individual transitions, HIV and Hep-C infection, crime, employment, and mortality. The costs and benefits was calculated based on individual life events. The initial state summary obtained yearly for 25 years of simulation model is illustrated in Fig. 2. The number of individuals who are incarcerated every year (for 25 years of simulation) and the type of crime is presented in Fig. 3. Costs such as state costs, crime costs, value of life year costs, HIV & Hep-C treatment costs, and family burden costs are presented in Fig. 4. The mortality, HIV cases, and individuals with over 60 years of age

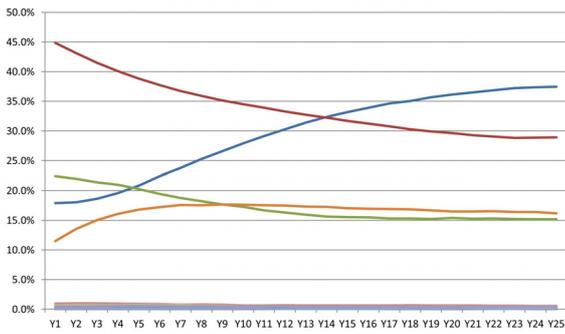


Fig. 2. State summary for 25 years of simulation

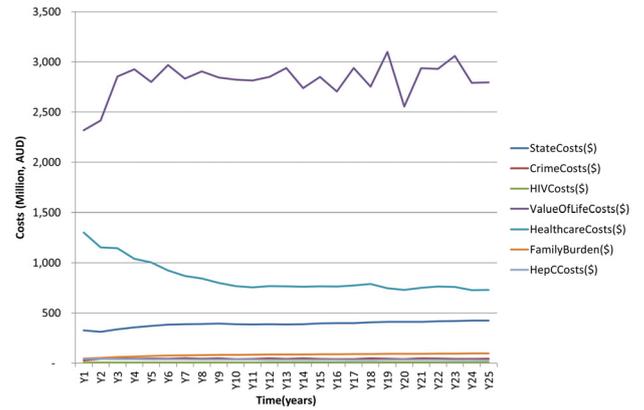


Fig. 4. Projected costs for 25 years of simulation

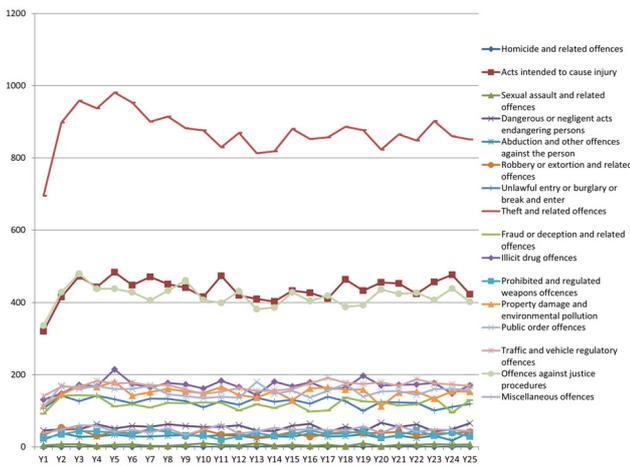


Fig. 3. Crime rates for 25 years of simulation

for 25 years of simulation is presented in Fig. 5. The current Heroin using population and their HepC status is presented in Fig. 6. The purpose of this paper was to showcase the model workings, initial validation and results. These results of the model can drive the discussion on government investments on drug treatment policy and decision making. It can also be used to simulate alternative policy scenarios, which is a topic of future research.

DISCUSSION AND CONCLUSION

The ISM model developed in this paper has modelled range of treatment services in community and prison stages, and used variety of datasets from illicit drug user surveys and published literature. Traditionally, variety of economic evaluations based on simulation models in the field of illicit drug use had been developed. Nevertheless, most of these existing models were cohort-based models and lacks consideration for individual histories and attributes. The major weakness of cohort-based approach was that the future individual events did not depend on prior events. Similar simplifications can lead to misleading model based estimations. Individual sampling models based on microsimulation are starting to be used in health-care decision-making. These models have better ability to represent heterogeneity that is required in complex modelling scenarios

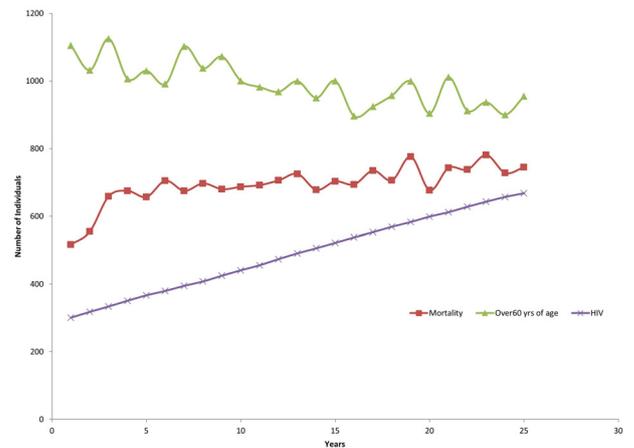


Fig. 5. Mortality, HIV status, and individuals over 60 yrs of age

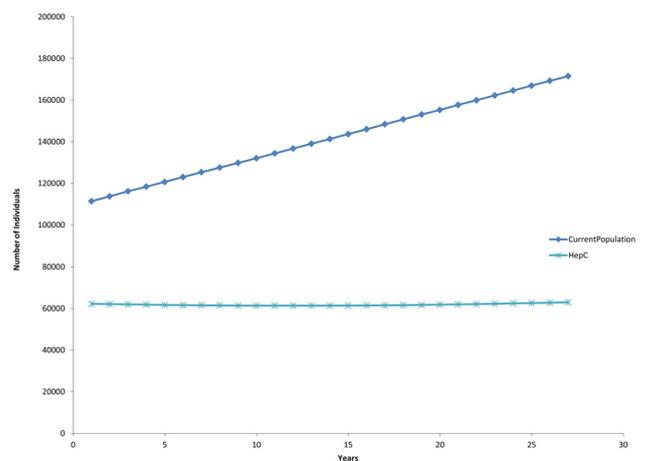


Fig. 6. Current Population of Heroin Users and their HepC status

such as illicit drug use. Individuals in these models can take into account their histories to make decision on their next transition without creating large number of health states. Various features of proposed model for heroin use were validated against external datasets and published data in this field, which were not used as an input for model parameterisation. The model output seems coherent and does not significantly diverge from external datasets.

ACKNOWLEDGMENT

This work is supported by Project Grant (APP1042923) from National Medical and Health Research Council (NMHRC), Australia.

REFERENCES

- World Bank. World Development Indicators 2012. *Washington DC*.
- Schackman, B. R., Leff, J. A., Polsky, D., Moore, B. A., & Fiellin, D. A. Cost-effectiveness of long-term outpatient buprenorphine-naloxone treatment for opioid dependence in primary care. *Journal of General Internal Medicine*, 27(6), 669–676, 2011.
- Feldstein, M.S. Net Social Benefit Calculation and the Public Investment Decision *Oxford Economic Papers*, 16, 1964.
- MBS 2015 Medicare Benefits Schedule (MBS) Accessed: 4/03/2015, <http://www.health.gov.au/internet/mbsonline/publishing.nsf>
- Degenhardt L, Gisev N, Trevena J, Larney S, Kimber J, Burns L, Shanahan M, Weatherburn D. Engagement with the criminal justice system among opioid-dependent people: a retrospective cohort study. *Addiction*. 2013 Dec;108(12):2152-65.
- Burns L, Randall D, Hall WD, Law M, Butler T, Bell J, Degenhardt L. Opioid agonist pharmacotherapy in New South Wales from 1985 to 2006: patient characteristics and patterns and predictors of treatment retention. *Addiction*. 2009 Aug;104(8):1363-72.
- Bell J, Mattick RP, Chan J, Hay A, Hall W. Methadone maintenance and drug related crime. *Journal of Substance Abuse Treatment* 9; 15-25, 1997.
- Coffin PO, Sullivan SD. Cost-effectiveness of distributing naloxone to heroin users for lay overdose reversal. *Ann Intern Med*. 158(1):1-9, 2013.
- Siebert U1, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, Kuntz KM. State-transition modeling: A report of the ISPOR-SMDM modeling good research practices task force-3. *Value in Health*, 15(6), 812-20, 2012.
- Hser YI, Huang D, Chou CP, Anglin MD. Trajectories of heroin addiction: growth mixture modeling results based on a 33-year follow-up study. *Eval Rev*. 31(6):548-63, 2007.
- Nosyk, B., Anglin, M.D., Brecht, M.L., Lima, V.D., Hser, Y.I. Characterizing durations of heroin abstinence in the California Civil Addict Program: results from a 33-year observational cohort study. *Am. J. Epidemiol.*, 177: 675682, 2013.
- Shah, N.G., Galai, N.G., Celentano, D.D., Vlahov, D., Strathdee, S.A. Longitudinal predictors of injection cessation and subsequent relapse among a cohort of injection drug users in Baltimore, MD, 1988-2000. *Drug Alcohol Depend.*, 83: 147156, 2006.
- Termorshuizen F, Krol A, Prins M, Van Ameijden E. J. C. Long-term outcome of chronic drug use the Amsterdam cohort study among drug users. *Am J Epidemiol* 161: 271279, 2005.
- Simpson, D.D. & Marsh, K.L. Relapse and recovery among opiate addicts 12 years after treatment. In *Tims, F.M. and Leukefeld, C.G. (Eds) Relapse and Recovery in Drug Abuse*. NIDA Research Monograph 72, 1986.
- Larney S & Indig D. Trends in illicit drug use in Australian correctional centres. *Annual meeting of the college of problems on drug dependence*. California: Palm Springs, 2012.
- Grnbladh, L., Gunne, L. Methadone-assisted rehabilitation of Swedish heroin addicts. *Drug and Alcohol Dependence*, Volume 24(1):3137, 1989.
- Hser Y I, Evans E, Huang D, Anglin D M. Relationship between drug treatment services, retention, and outcomes. *Psychiatr Serv*;55(7):767-74.
- Zhang, Zhiwei, Friedmann, Peter D., Gerstein, Dean R. Does retention matter? Treatment duration and improvement in drug use. *Addiction*, 98(5): 673-684, 2003.
- Zucchelli, E., Jones, AM and Rice, N. The evaluation of health policies through dynamic microsimulation methods. *International Journal of Microsimulation*, 5(1), 2-20.
- Russell G Smith, Penny Jorna, Josh Sweeney, Georgina Fuller. Counting the costs of crime in Australia: A 2011 estimate. *AIC Reports Research and Public Policy Series* 129, ISBN 978 1 922009 70 8 ISSN 1836-2079, 2013.

OPTIMAL FISCAL POLICIES AFTER THE “GREAT RECESSION”: A CASE STUDY FOR SLOVENIA

Reinhard Neck and Dmitri Blueschke
Department of Economics
Alpen-Adria-Universität Klagenfurt
A-9020 Klagenfurt, Austria
reinhard.neck@aau.at, dmitri.blueschke@aau.at

Klaus Weyerstrass
Macroeconomics and Public Finances
Institute for Advanced Studies
A-1080 Vienna, Austria
klaus.weyerstrass@ihs.ac.at

KEYWORDS

Macroeconomics; fiscal policy; optimization; optimal control; Slovenia; crisis; public debt.

ABSTRACT

In this paper, we investigate how fiscal policies should look like in a country like Slovenia. Slovenia's present situation is characterized by high and rapidly increasing public debt and low growth. It is an interesting case because it is one of the few small open economies from Central and Eastern Europe that was already in the Euro Area before the “Great Recession”. Using the SLOPOL model, an econometric model of the Slovenian economy, we analyse the effects of different fiscal policies in Slovenia over the next couple of years by means of simulations. In particular, we determine optimal fiscal policies for Slovenia over the next few years. Using the OPTCON algorithm, we calculate approximately optimal fiscal policies under different scenarios. We show that an optimal design of fiscal policies depends essentially on the political preferences of Slovenian policy makers. Moreover, the simulations and optimizations reveal the small scope of possible alternative fiscal stabilization policies available due to the relatively low effectiveness of the fiscal instruments with respect to their influence on the business cycle in the Slovenian economy.

INTRODUCTION

The “Great Recession”, the financial and economic crisis of 2007 to 2009, was the most severe economic crisis since the Great Depression of the 1930s. It resulted in negative growth and increasing unemployment in nearly all industrial countries, irrespective of their initial situation. Some countries, however, suffered particularly hard; this was partly due to government failures, i.e. to inadequate actions taken by their economic policy makers. This is also true for Slovenia, whose economic situation deteriorated very badly until 2014. Together with most former socialist countries from Central and Eastern Europe, Slovenia entered the European Union in 2004, and it managed to introduce the euro as legal tender as early as 2007. The economic development of Slovenia

was successful in terms of GDP growth and the decrease in unemployment before the “Great Recession”.

However, this positive macroeconomic development disguised a housing bubble. With the outbreak of the global financial and economic crisis, the real estate bubble burst, and the impact of the recession was especially deep in Slovenia, with a decline in GDP of almost 8 percent in one single year (2009) and an increase in unemployment to the same level as in the year before Slovenia joined the Euro Area (2006). In the following years, the country was hit especially hard by another financial and economic crisis, which, in addition, resulted in an unprecedented increase in its government debt. The economic crisis culminated in a severe financial crisis in 2013. This required significant public support to six banks, at a fiscal cost of about 10 percent of GDP. As a result, Slovenia's fiscal position deteriorated significantly. The budget deficit rose from near zero in 2007-2008 to almost 14 percent of GDP in 2013, and the debt ratio quadrupled, rising to 80 percent in 2014. After three years of low or even negative growth, first signs of economic stabilization became visible in 2014 with rising exports and investment.

Although by now a lot of evidence on the effects of macroeconomic policies during the “Great Recession” is available, its interpretation still diverges among members of different economic schools. In particular, the role of fiscal policy and the specific problems of countries in the Euro Area are subject to ongoing controversies. In general, fiscal policy effects are smaller *ceteris paribus* in small open economies than in larger economies that are less exposed to global shocks. Furthermore, a high level of public debt is likely to undermine positive effects of fiscal stimuli. Hence, a clear commitment to fiscal consolidation after overcoming a crisis is required. However, strict fiscal consolidation measures in a recession may contribute to a deepening of the recession (Blanchard and Leigh 2013).

In this paper, we analyse the effects of different fiscal policy scenarios in Slovenia over the next couple of years and evaluate them according to their effects on macroeconomic target variables. We use the SLOPOL model, an econometric model of the Slovenian economy constructed by us to make forecasts and simulate the effects of the global and the European crisis under alternative assumptions. In particular, we consider the assumption of no-policy reactions, i.e. assuming that fiscal policies do not attempt to deal with the effects of

the crisis. Moreover, we determine optimal fiscal policies for Slovenia over the next few years. We use the SLOPOL model and assume an intertemporal objective function for Slovenian policy makers containing output, unemployment, inflation, the budget deficit, public debt and the current account as arguments. Using the OPTCON algorithm, we calculate approximately optimal policies under different scenarios. It turns out that there are some trade-offs between the design of countercyclical fiscal policies and the requirements of fiscal solvency. The resulting optimal fiscal policies are only mildly countercyclical and can protect the Slovenian economy from negative effects of a recession only in a very limited way.

THE MACROECONOMETRIC MODEL SLOPOL8

For this study, we use an updated version of the SLOPOL (SLOvenian POLicy model) model. SLOPOL is a medium-sized macroeconomic model of the small open economy of Slovenia. We use the version SLOPOL8 consisting of 61 equations, of which 24 are behavioural equations and 37 are identities. In addition to the 61 endogenous variables that are determined in the equations, the model contains 29 exogenous variables. For the present work, we updated the SLOPOL8 version as described in Blueschke et al. (2016) until the end of 2015. The behavioural equations were estimated by ordinary least squares (OLS), and most of them were specified in error correction form.

The model contains behavioural equations and identities for several markets and sectors: the goods market, the labour market, the foreign exchange market, the money market and the government sector. Rigidities of wages and prices are present. The model combines Keynesian and neoclassical elements, the former determining the short and medium run solutions in the sense that the model is demand-driven and persistent disequilibria in the goods and labour markets are possible.

The supply side incorporates neoclassical features. Potential output is determined by a Cobb-Douglas production function with constant returns to scale. It depends on trend employment, the capital stock and autonomous technical progress. Trend employment is equal to the labour force minus natural unemployment, the latter being defined via the non-accelerating inflation rate of unemployment (NAIRU). The NAIRU, which approximates structural unemployment, is estimated by applying the Hodrick-Prescott filter to the actual unemployment rate. For forecasts and simulations, the structural unemployment rate is then extrapolated with an autoregressive process. For forecasts, technical progress is extrapolated exogenously.

On the demand side, private consumption depends on current disposable income and on lagged consumption. In addition, the long-term real interest rate enters the consumption equation with a negative sign. Real gross fixed capital formation depends on the change in total domestic demand (in accordance with the accelerator hypothesis) and by the user cost of capital, where the

latter is equal to the real interest rate plus the depreciation rate of the capital stock. Changes in inventories are exogenous in the SLOPOL model.

Real exports of goods and services are a function of the real exchange rate and of foreign demand for Slovenian goods and services, where for the latter we use the volume of world trade as a proxy. Real imports of goods and services depend on domestic final demand and on the real exchange rate.

On the money market, the short-term interest rate depends on its Euro Area counterpart to capture Slovenia's Euro Area membership and the resulting gradual adjustment of interest rates in Slovenia towards the Euro Area average. In the same vein, the long-term Euro Area interest rate is included in the equation determining the long-term interest rate in Slovenia. In addition, the long-term interest rate depends on the short-term rate, representing the term structure of interest rates. Due to Slovenia's membership of the Euro Area, the nominal exchange rate is exogenous for Slovenia. However, the real exchange rate is still endogenous even for the Euro Area countries, since it also depends on the domestic price development. The bilateral exchange rate between the Slovenian tolar and the euro is included as one of the explanatory variables in the real effective exchange rate equation. In addition, the exchange rate between the euro and the US dollar and the Slovenian rate of inflation are also regressors in this equation.

The labour demand of companies (actual employment) depends on the final demand for goods and services as well as on unit labour costs, the latter being equal to the nominal gross wage divided by labour productivity. Labour productivity in turn is equal to real GDP per employee. Labour supply by private households is determined by the participation rate, i.e. the labour force (employed plus unemployed persons) divided by the working-age population (the population aged 15 to 64). The participation rate depends on the real net wage.

In the wage-price system, gross wages, the CPI and various deflators are determined. The gross wage rate depends on the price level, labour productivity and the difference between the actual and the natural rate of unemployment (or the NAIRU). Consumer prices depend on domestic and international factors. The former comprise unit labour costs and the capacity utilisation rate. In addition, Slovenian prices depend on the oil price, converted into domestic currency. The GDP deflator and the deflators for private and public consumption are linked to consumer prices. The export deflator depends on unit labour costs in Slovenia and on world trade. Finally, the import deflator is influenced by the oil price in euros as a proxy for international raw material prices, which constitute an important determinant of the price level in a small open economy like Slovenia.

In the government sector of the model, the most important expenditure and revenue items of the Slovenian budget are determined. Social security contributions by employees are equal to the product of the average social security contribution rate, the gross wage rate and the number of employees. In the same vein,

income tax payments by employees are the product of the average income tax rate, the gross wage rate and the number of employees. In a behavioural equation, social security payments by companies depend on social security contributions by employees. Profit tax payments by companies depend on nominal GDP as an indicator for the economic situation. Value added tax revenues depend on the value added tax rate and on private consumption. Finally, the remaining government revenues are explained by nominal GDP, considering the fact that they are also pro-cyclical.

On the expenditure side of the budget, interest payments depend on the stock of public debt and on the long-term interest rate. Finally, the remaining government expenditures are, as in the case of the revenues, determined by nominal GDP as an indicator of the economic situation. The budget balance is equal to the difference between total government revenues and expenditures. The public debt level is extrapolated using the budget balance equation.

SIMULATION EXPERIMENTS

We perform simulations over five years, i.e. the period 2016 to 2020. For ex ante simulations, i.e. simulations for a period in the future, we need to specify plausible paths for the exogenous variables and policy instruments. Using these time paths, we perform a baseline run. Then alternative scenarios with more active fiscal policy can be defined and compared to the baseline simulation. Among the “truly” exogenous variables, i.e. those beyond the control of domestic policy makers are principally all international variables. In the version of

the Slovenian model used here, these comprise world trade, the exchange rate between the euro and the US dollar, the short-term interest rate, which is mainly determined by the European Central Bank, and the average long-term interest rate, i.e. the yield on 10-year government bonds, in the Euro Area. There are reasons to assume that in the future world trade will grow less than before the outbreak of the crisis: in some emerging economies the growth trend has slowed down, the integration of the emerging economies in the world economy has already progressed considerably, and many industrialised economies are still struggling with deleveraging in the public or private sector.

In accordance with the world trade volume, we assume the Euro Area short- and long-term interest rates to normalize in the coming years. More specifically, we assume that the 3-month Euribor and the average Euro Area long-term interest rate will rise to 2 percent and 3 percent in 2020 respectively. For the fiscal policy instruments, we assume that government consumption, public investment and transfer payments to private households increase by 4 percent per year over the entire simulation period. The average personal income tax rate remains constant at its 2015 level.

Table 1 shows the results of the baseline simulation. Due to the assumed slowdown of world trade, real GDP growth, which is still high in 2015, decreases gradually over the simulation period. This development results in only a slight increase of unemployment, which then decreases towards pre-recession levels. The labour force declines because of a decline of the population of working age. Another reason is the stagnation of the real wage due to the weak demand for Slovenian exports and to a small increase in inflation.

Table 1: Budget Consolidation Strategies – Baseline

	2015	2016	2017	2018	2019	2020
GDP growth (%)	2.9	2.0	1.5	0.8	1.0	0.8
Inflation (%)	-0.8	0.6	1.3	1.3	1.1	1.0
Unemployment rate (%)	11.7	12.0	10.4	9.9	9.6	9.6
Budget balance ratio to GDP (%)	-3.3	-2.4	-2.6	-3.2	-3.6	-4.2
Debt ratio to GDP (%)	73.3	75.0	76.7	79.2	81.8	85.1

Due to the unfavourable development of the real economy, the budget deficit is gradually increased. In 2018, the Slovenian budget deficit ratio surpasses the 3 percent reference level laid down in the Stability and Growth Pact (SGP). Accordingly, the debt ratio rises further from 73 percent in 2015 to 85 percent in 2020.

Beside the baseline scenario as presented above, we ran several alternative scenarios with more active fiscal policy. Lack of space precludes a detailed discussion of them. Cf. also the results obtained with an earlier version of the model in Blueschke et al. (2016). Their main result

is a weak effect of fiscal policy measures on the real economy, in particular on the labour market. In addition, the link between budgetary developments and the growth of GDP is weaker than expected for a model mainly following Keynesian lines.

OPTIMIZATION EXPERIMENTS

Simulations of alternative economic developments under varying assumptions about economic policy measures are the main instrument for empirical analyses of

macroeconomic policy with econometric models. However, they suffer from the arbitrary character of the assumptions made about the policy instruments and the lack of a systematic choice of scenarios. Alternatively, one might determine “optimal” policies using the optimum control framework. Under this approach, it is necessary to formulate an objective (or loss) function summarizing the time paths of the different objective variables (instruments and endogenous target variables) into one scalar to be optimized (maximized or minimized) by a (hypothetical) policy maker.

Optimum control theory provides the mathematical tools for obtaining optimal policy trajectories when using a dynamic econometric model such as SLOPOL8. As usual in economic policy applications, we assume a quadratic intertemporal objective function involving deviations of the values of the relevant variables from some pre-specified “ideal” paths. The objective function has the following form:

$$L = \frac{1}{2} \sum_{t=1}^T \begin{bmatrix} \mathbf{x}_t - \tilde{\mathbf{x}}_t \\ \mathbf{u}_t - \tilde{\mathbf{u}}_t \end{bmatrix} \mathbf{W}_t \begin{bmatrix} \mathbf{x}_t - \tilde{\mathbf{x}}_t \\ \mathbf{u}_t - \tilde{\mathbf{u}}_t \end{bmatrix},$$

$$\mathbf{W}_t = \alpha^{t-1} \mathbf{W}, t = 1, \dots, T.$$

\mathbf{x}_t denotes the vector of state variables and \mathbf{u}_t the vector of control variables, $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{u}}_t$ are the desired (“ideal”) values of the state and control variables, \mathbf{W} is the matrix of the weights given to the deviations of the state and control variables from their respective desired values, and α denotes the discount factor. The policy maker aims at minimizing this objective function subject to the set of dynamic constraints given by the econometric model. With the nonlinear econometric model SLOPOL8, this results in a multivariable nonlinear-quadratic optimal control problem. An exact solution to such a problem is not possible, so we have to resort to numerical approximations. Here we use the OPTCON2 algorithm; cf. Matulka and Neck (1992), Blueschke-Nikolaeva et al. (2012) for details.

Although this algorithm allows for a rather elaborate menu of stochastic extensions, here we confine ourselves to deterministic optimal control, assuming the model parameters and the model equations to be exactly true. Apart from the considerable reduction in computing time achieved by this simplification, the main reasons for it are the limited amount of reliable information about the stochastics of the model and our experience that stochastic control results often come close to deterministic ones.

The policy maker in this optimal control experiment is the government of Slovenia. We assume that it has three control variables at its disposal: government consumption (GN), transfers (TRNSFERSN) and government investments (GINVN). We select seven state variables for which we define certain “ideal” paths and which enter the objective function. These are: the growth rate of GDP (GRGDPR), the level of real GDP (GDPR), the unemployment rate (UR), the inflation rate (INFL), the budget balance ratio to GDP (BALANCEGDP), the ratio of the debt level to GDP (DEBTGDP) and the ratio

of the current account balance to GDP (CAGDP). By this choice of targets, we aim at representing the most important goals of macroeconomic policy making. The “ideal” paths imply smooth growth in the income variables and low values for the rates of unemployment and inflation. In more detail, “ideal” GDPR grows by 3%; hence, the “ideal” value of GRGDPR, i.e. the growth rate of real GDP, is set to 3%. The “ideal” inflation rate is set to be 2%, which is in accordance with the declared target of the ECB. “Ideal” values for CAGDP and BALANCEGDP are set to zero. The “ideal” debt level is required to decrease from 74% (2015) to the Maastricht threshold of 60% of GDP (2020). The “ideal” unemployment rate should decrease in a linear way, starting with its current value of around 12% (in 2015) and arriving at a value of around 6% at the end of the planning horizon (in 2020).

For this paper, we ran four optimum control experiments, called “EQUALW”, “DEBT”, “GRGDPR” and “GDPR”. The differences between the experiments arise from different weights for the objective variables. In other words, the different scenarios give different importance to certain objective variables. Table 2 summarizes these weights.

In the EQUALW optimization scenario, all seven objective state variables get the same “raw” weights. These raw weights have to be normalized according to the characteristics of the corresponding time series; see Blueschke (2014) for more details. In the DEBT scenario, we give a higher weight to public debt, which should deliver the optimal results for a more severe fiscal consolidation policy. In the GRGDPR scenario, we give a higher weight to the growth rate of GDP, which should give the optimal results for a growth-oriented fiscal policy strategy. The idea behind the GDPR scenario is quite similar, namely giving a high weight to the level of real GDP. The main difference is that this strategy will be less discretionary than the GRGDPR scenario.

Table 2: Raw Weights of the Objective Variables for Four Optimal Control Scenarios

Weight	EQUALW	DEBT	GRGDPR	GDPR
CAGDP	1	1	1	1
GRGDPR	1	1	10	1
UR	1	1	1	1
INFL	1	1	1	1
GDPR	1	1	1	10
BALANCEGDP	1	1	1	1
DEBTGDP	1	10	1	1

The resulting values of the objective function (which is minimized) are reported in Table 3. Here “J_simulation” denotes the objective function value of the non-controlled simulation (the baseline simulation from the previous section, with the weights in the objective function set according to the respective column of Table 2. “J_OPTCON” denotes the objective function value for the optimal control strategy, again resulting from the values of the weights given in Table 2.

Table 3: Values of the objective function for the optimal control experiments

	EQUALW	DEBT	GRGDPR	GDPR
J_simulation	218.06	582.66	685.76	403.43
J_OPTCON	209.39	328.18	571.70	385.41

The results show that in all experiments, the optimization improves the system performance over the uncontrolled simulation results and arrives at considerably lower values of the objective function. We can also see from Table 3 that especially the DEBT experiment brings about a huge improvement: the optimal solution's loss is lower than the non-controlled solution's loss by 43.7%. The worst performance, both without and with optimization, results from the GRGDPR scenario, which shows the high costs of attempts to achieve point targets for growth rates in a discretionary way, especially in an economy like the Slovenian, which strongly depends on global developments. In contrast, in the EQUALW and in the GDPR scenario, the improvements are significantly smaller, which indicates the low effectiveness of fiscal policy and gives some additional support for a fiscal consolidation policy in the present situation in Slovenia.

Figures 1 to 9 show the resulting time paths of the main variables of the model in the scenarios described. Although the model is a quarterly one, we only show annual results. The quarterly time paths exhibit strong seasonal patterns (as do the data), which are irrelevant from the point of view of policy making.

The EQUALW optimization scenario requires more fiscal consolidation than the non-controlled simulation in all periods except the last one (2020), where the economic situation finally stabilizes sufficiently to phase out the austerity policy. In the first year of the optimization period (2016), this fiscal consolidation strategy leads to a growth rate of GDP that is close to 0.3 percentage points below the non-controlled solution. In the following years of the planning horizon, the growth rate of GDP in the EQUALW solution matches that of the non-controlled solution. Because of the lower growth rate of GDP, the EQUALW solution also leads to slightly (by less than 0.1 percentage points) higher unemployment rates than in the non-controlled simulation. The trade-off between economic growth and fiscal consolidation is very weak in the short run (and a fortiori in the medium and long run). Hence, the Slovenian economy has to accept only a tiny burden on the real economy in the next few years if the aim is to set public finances straight. However, as can be seen from Figures 8 and 9, the EQUALW solution does not allow consolidating public finances in Slovenia. The budget deficit misses the 3% Maastricht threshold in the last two years, and public debt violates the 60% threshold throughout the planning period. Only a modest improvement occurs over the non-controlled simulation, with public debt increasing to 82.6% of GDP instead of 85% of GDP in the latter.

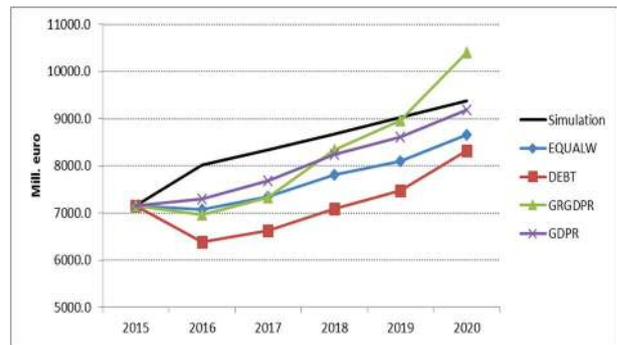


Figure 1: Government Consumption

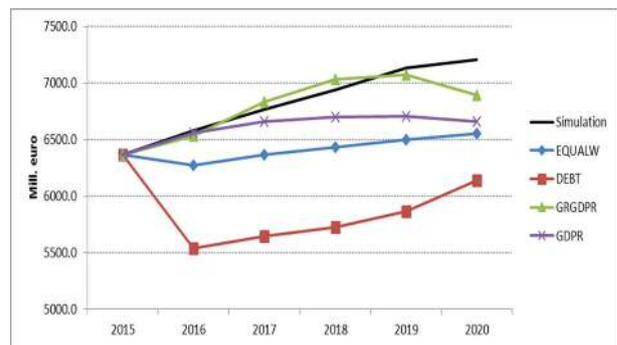


Figure 2: Transfers

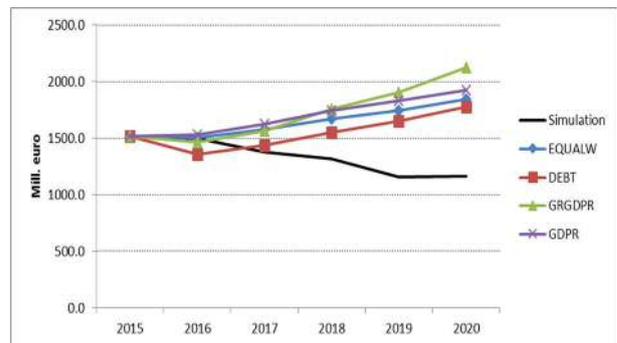


Figure 3: Government investment

The DEBT scenario shows optimization results for a hypothetical case in which the government of Slovenia places high importance on the target of decreasing public debt (or, rather, its ratio to GDP). In such a case, a much more strongly restrictive fiscal policy is optimal. Government consumption decreases by 10.6%, transfers by 13.1% and government investment by 10.5% in the first year of the optimization horizon (2016). It is noteworthy that the fiscal restriction affects transfers most and public investment less, which accords with policy recommendations for austerity policies derived from models including also supply side considerations. This strong intervention leads to a drop in the growth rate of GDP by nearly one percentage point and initiates a period of slow (around one percent) growth during the following years for the Slovenian economy. The effect is politically relevant as the level of GDP is significantly below the uncontrolled simulation and does not recover

over time. This policy produces higher rates of unemployment, which are, however, only less than one percentage point above the non-controlled and the EQUALW solutions. The rate of inflation is lower than in the other solutions.

The very restrictive fiscal policy leads to budget surpluses over three years, but even in this scenario it turns into negative later on and misses the Maastricht criterion in the last year. The latter effect may be due to the neglect of future periods beyond 2020, which is unavoidable given the limitations of the present version of OPTCON. For a long-run solution, one would have to consider convergence to a steady state of the dynamic system with a balanced budget, which would require a more thorough investigation of the supply side of the Slovenian economy and a different econometric model. Here public debt stabilizes in the sense of remaining within a band between 69 and 72 percent of GDP over the entire planning horizon and arrives at 71.9% of GDP in 2020. This still violates the Maastricht and SGP criterion but represents a non-negligible improvement compared to the other scenarios.

The GRGDPR and GDPR scenarios attach great importance to output (and implicitly to employment). The difference between these two scenarios is that the GRGDPR scenario focuses on the growth rate of GDP and the GDPR scenario on the level of GDP.

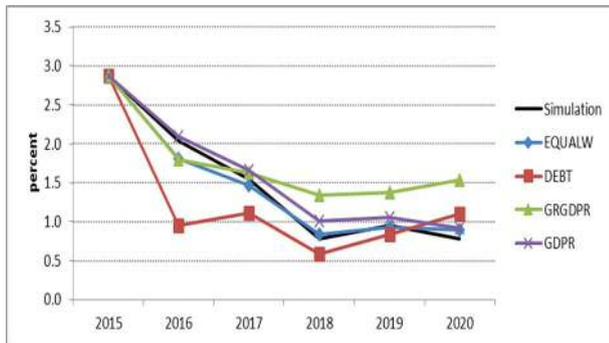


Figure 4: Real GDP Growth Rate

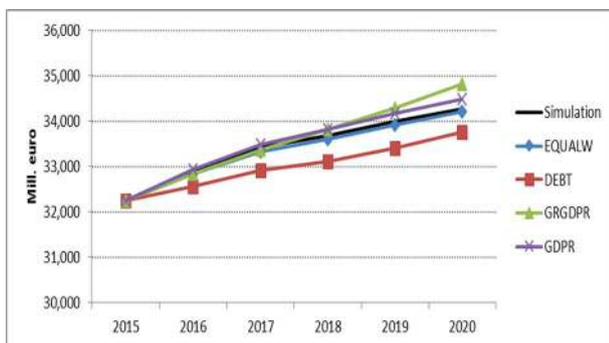


Figure 5: Real GDP Level

Both scenarios require a less restrictive fiscal policy than the EQUALW scenario. As expected, the GRGDPR scenario seems to be more discretionary, with more active expansionary fiscal intervention in the later years

especially in terms of the government consumption and government investment control variables (and partly at the expense of transfers toward the end of the planning horizon). Because of this output-oriented policy, both scenarios lead to slightly higher growth rates of GDP and levels of GDP and to lower values for the unemployment rate than in the EQUALW solution. This effect increases over time due to the weak growth prospects of the Slovenian economy implied by the baseline simulation. However, these improvements in the real economy are minimal, which demonstrates the low effectiveness of fiscal policy with respect to real variables. This result is primarily due to the fact that the economy of Slovenia is mainly influenced by the global economic development. This is in line with the current literature on globalization issues, which shows the increasing influence of global factors, especially for small open economies such as Slovenia.

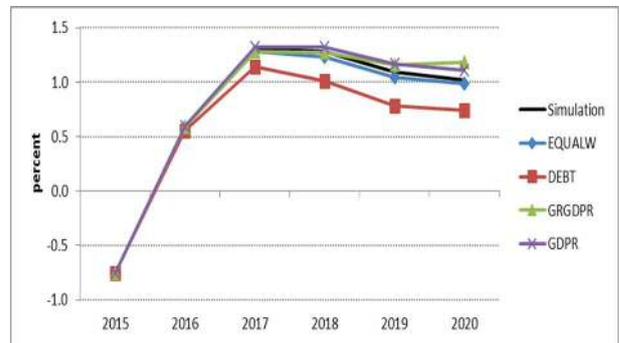


Figure 6: Inflation Rate

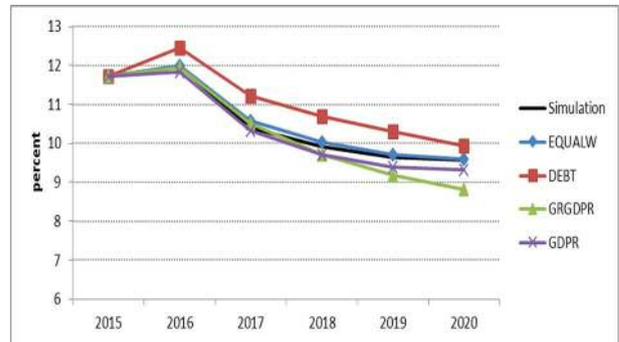


Figure 7: Unemployment Rate

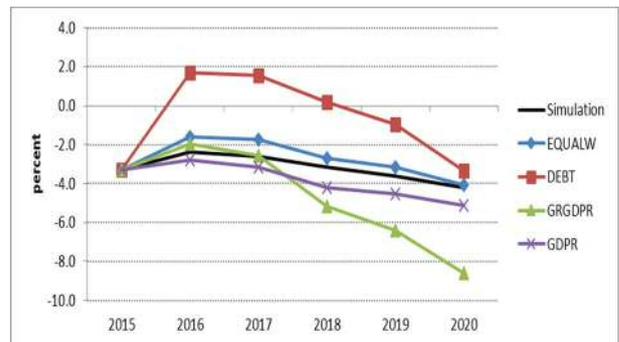


Figure 8: Budget Balance Ratio to GDP

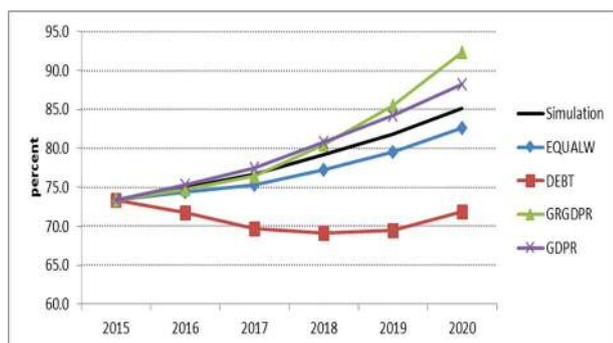


Figure 9: Public Debt Ratio to GDP

CONCLUSIONS

In the present study, we analyse four optimal control scenarios for the Slovenian economy relating to its current government debt crisis. For this purpose, we use the SLOPOL8 model and assume a planning horizon of 5 years (2016-2020). Using the optimal control framework and the OPTCON algorithm, we run four different optimization experiments. In the so-called EQUALW optimization experiment, we give equal importance to all objective variables. In the remaining scenarios, we place high importance on either public debt or output (growth or level). This optimum control exercise provides several insights such as identifying various trade-offs. In the EQUALW scenario, a very mild fiscal consolidation policy turns out as optimal.

The debt-oriented optimization scenario results in the greatest improvement in the objective function value with respect to the uncontrolled solution and requires a restrictive fiscal policy especially in the first year of the optimization period, which we interpret as an investment into an increased room for manoeuvring in the future. The effects of such a policy are, as to be expected, negative for the real economy, but involve only transitory costs in terms of unemployment. The growth rates of GDP come close to those in the EQUALW and in the output-oriented scenarios in the last period (2020). In the scenarios with great importance placed on output, the optimal results are more expansionary than in the simulation with no policy intervention and in the EQUALW optimization. In general, fiscal policy has a relatively small effect on the output objectives, and even increasing the importance of these targets by a factor of 10 is not sufficient for a highly expansionary fiscal policy to be optimal. These results indicate the low effectiveness and high costs of an expansionary use of fiscal instruments in a small open economy like Slovenia.

Of course, it would be premature to infer strong conclusions for the current macroeconomic situation of the Slovenian economy based on just one model specification and a few optimization runs, but our results may lend some support to an austerity course as recommended, for instance, in the Slovenian Stability Programme.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the Austrian Science Fund (FWF): project I 2764-G27. Thanks are due to Viktoria Blueschke-Nikolaeva for collaboration on the OPTCON algorithm and to three anonymous referees for helpful suggestions. The usual caveat applies.

REFERENCES

- Blanchard, O. and D. Leigh. 2013. "Growth Forecast Errors and Fiscal Multipliers." *American Economic Review*, 103, May, 117-120.
- Blueschke, D. 2014. *Optimal Policies for Nonlinear Economic Models: The OPTGAME3 and OPTCON2 Algorithms: Theory and Applications*. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken.
- Blueschke, D.; K. Weyerstrass; and R. Neck. 2016. "How Should Slovenia Design Fiscal Policies in the Government Debt Crisis?" Forthcoming in *Emerging Markets Finance and Trade*.
- Blueschke-Nikolaeva V.; D. Blueschke; and R. Neck. 2012. "Optimal Control of Nonlinear Dynamic Econometric Models: An Algorithm and an Application." *Computational Statistics and Data Analysis* 56, 3230-3240.
- Coenen, G.; M. Mohr; and R. Straub. 2008. "Fiscal Consolidation in the Euro Area: Long-run Benefits and Short-run Costs." *Economic Modelling* 25, 912-932.
- Cogan J.F.; T. Cwik; J.B. Taylor; and V. Wieland. 2010. "New Keynesian versus Old Keynesian Government Spending Multipliers." *Journal of Economic Dynamics and Control* 34, 281-295.
- Matulka, J. and R. Neck. 1992. "OPTCON: An Algorithm for the Optimal Control of Nonlinear Stochastic Models". *Annals of Operations Research* 37, 375-401.

AUTHOR BIOGRAPHIES

REINHARD NECK was born and studied in Vienna. After having held positions at universities in Switzerland, Germany and the USA, he became and is now Professor of Economics in Klagenfurt, Austria. His e-mail address is: reinhard.neck@aau.at.

DMITRI BLUESCHKE was born in Kazakhstan. He studied economics at the University of Bielefeld until 2007. He got his Ph.D. at the Alpen-Adria-Universität Klagenfurt in 2011, where he is now Assistant Professor. His e-mail address is: dmitri.blueschke@aau.at.

KLAUS WEYERSTRASS was born in Germany and studied at the University of Osnabrück. He got his PhD at the Alpen-Adria-Universität Klagenfurt in 2000 and is now with the Institute for Advanced Studies in Vienna. His e-mail address is: klaus.weyerstrass@ihs.ac.at.

FRIENDSHIP OF STOCK INDICES

László Nagy

Mihály Ormos

Department of Finance

Budapest University of Technology and Economics

Magyar tudósok körútja 2, Budapest H-1117, Hungary

E-mail: nagy1@finance.bme.hu, ormos@finance.bme.hu

KEYWORDS

cluster analysis, equity index networks, machine learning

ABSTRACT

The aim of this study is to cluster different stock indices based on historical time series data.

The current research shows that tail events have minor effect on the equity index structure. It also turns out that major part of the total variance can be explained by clusters. In addition, clusterwise regressions are reliable, hence CAPM with clusters gives real information about risk and reward.

INTRODUCTION

The average surface temperature of the Earth is 15 °C. Everybody feels the temperature; however, it does not say too much about the current local conditions. Seasonal and geographical adjustments are required. Similarly, the global stock market structure has to be well understood to analyse local economic trends. Institutional economic surveys mostly provide qualitatively identified network structures e.g. emerging markets, developed markets.

The main goal of this study is to provide quantitative techniques to discover the equity index network structure.

The baseline concept follows the CAPM, in which similarity measures are calculated from correlations between logarithmic returns (Yalamova 2009). The anomalies of CAPM indicate a two dimensional mean-beta framework that gives only a simplified picture of the real market structure. The proposed non-linear similarity kernels are able to deal with higher order terms, hence clusters would be more accurate.

We show that normalized Laplacian based spectral clustering techniques can be used for recognizing well separated clusters in the global financial markets. Analysing the correlation structure of stock indices turns out clusters are homogeneously connected with each other, hence the normalized Newman-Girvan modularity matrix brings better clustering results (Bolla 2013).

DATA

The current study presents detailed analysis of 59 stock indices. We apply USD denominated stock splits and

dividends adjusted daily closing prices between 26/9/1990 and 21/9/2015. Data is provided by Thomson Reuters.

In order to underline the highly different characteristics of individual stock indices we present some monthly descriptive statistics.

Table 1: Descriptive statistics of monthly returns

Index	Mean	Variance	Skewness
.CSI300	0.018	0.056	-0.336
.XU100	0	0.026	-0.809
.DJI	0.012	0.009	-0.819
.UAX	-0.034	0.037	-0.721
.WORLD	0.004	0.002	-1.889

Notes: Table 1 shows the descriptive statistics of the monthly returns, where CSI300, XU100, DJI, UAX, WORLD represent the Shanghai Composite 300, Brose Istanbul 100, Dow Jones, Ukraine UX index and MSCI World index respectively.

Our selection criterion for covered stock indices is based on their classification in IMF Economic Outlook 2015 and the MSCI WORLD Index composition in 2015. In our analysis we allocate approximately the same weight to each region. Although, the number of countries is not equal in each region and the market capitalization differs as well, we rebalance the sample by choosing approximately ten indices from each IMF group.

We are also interested in the role of well diversified indices e.g. MSCI WORLD and EURO STOXX600, hence we put them into the analysis.

SPECTRAL CLUSTERING

In the 20th century, the appearance of large, complex data sets brought new challenges to develop methods which can be used to understand complicated structures. Spectral clustering techniques provide optimal, lower dimensional representation of multidimensional data sets. The idea is to represent the data structure as a weighted graph, and cut the graph along the different clusters. This approach leads to penalized cut optimization problems. Linear algebra and cluster analysis give powerful methods to find the optimal representations and minimized cuts.

Similarity matrix

If we would like to cluster different items, first the measurement of similarity has to be decided. In this study similarity of two stock indices (i, j) will be denoted by $W_{i,j}$. The goal is to penal differences and reward similarities. Logarithmic returns are easy to handle and keep all the information about the price processes.

$$r_i(t) = \ln\left(\frac{S_i(t)}{S_i(t-1)}\right) \quad (1)$$

where $S_i(t)$ represents the price of index i . The current study analyses multiple similarity approaches.

First, the Markowitz based squared correlation is considered as a similarity metric.

$$W_{i,j} = \text{Corr}^2(r_i, r_j) \quad (2)$$

We argue this approach because logarithmic returns are not normally distributed, hence non-linear effects also could be important. However; correlation is linear, hence squared correlation similarities take into account only linear dependences.

The problem of higher-order moments can easily be solved by using symmetric and positive-definite kernel functions. The idea comes from the functional analysis. Data can be transformed into a reproducing kernel Hilbert space (RKHS) where applying the usual statistics provide the same outcomes which can be reached by using non-linear statistics in the original Hilbert space. In practice, the Gaussian-kernel is widely used (Leibon et al. 2008).

$$W_{i,j} = \exp(-\|r_i - r_j\|^2) \quad (3)$$

We notice that, if the sets of the relevant information and sensitivities are similar, then the relative entropy of the distribution of return processes is small. Otherwise, we can say stock indices are sensitive to different sets of information in a different manner (Ormos and Zibriczyk 2014). This means similarity function has to be monotonically decreasing in symmetric Kullback-Leibler distance, thus we can construct a similarity measure such that:

$$W_{i,j} = \frac{1}{1 + [KL(p(r_i)\|p(r_j)) + KL(p(r_j)\|p(r_i))]/2} \quad (4)$$

where $p(r_i)$ denotes the probability distribution function of logarithmic returns of index i and $KL(p(r_i)\|p(r_j)) \stackrel{\text{def}}{=} \sum_x p(r_i = x) \ln\left(\frac{p(r_i=x)}{p(r_j=x)}\right)$ is the relative entropy of indices i and j .

Another perspective says that large deviations are riskier, hence similarities should be defined with tail distributions. We calculate the differences of return series and count the number of at least two standard deviation peaks. The logic implies indices are similar if their price processes jump together. Similarity function

has to be decreasing in the number of large deviations, hence we propose the following metric;

$$W_{i,j} = \frac{1}{1 + \sum_{t=1}^T \delta(|z_i(t) - z_j(t)| > 2)} \quad (5)$$

where z_i represents the normalized return of index i . In the current study we compare each approaches.

Normalized modularity

The equity index structure is strongly connected. We can not say that events in Africa do not have any kind of effects on European markets, hence we have to find methods which can be used to cluster dense graphs.

Let $G(V_{N \times 1}, W_{N \times N})$ be a weighted graph, where V denotes the set of vertices and W represents the weights of the edges. A k -partition of graph $G(V, W)$ can be defined as the partition of vertices such that $\bigcup_{a=1}^k V_a = V$ and $V_i \cap V_j = \emptyset \quad \forall i, j \in \{1, \dots, k\}$.

The $W_{i,j}$ value represents the strength of the connection between nodes (i, j). If we assume that nodes are independently connected, then the guess of weight $W_{i,j}$ will be the product of the average connection strength of i and j . The average connection strength d_i and d_j are given by W ,

$$d_i = \frac{1}{N} \sum_{u=1}^N W_{i,u}$$

Thus, $W_{i,j} - d_i d_j$ captures the information of the network structure (Bolla 2011), hence if we would like to maximize the sum of information in each cluster, then we get:

$$\max_{P_k \in \mathcal{P}_k} \sum_{a=1}^k \sum_{i,j \in V_a} (W_{i,j} - d_i d_j) \quad (6)$$

where P_k stands for specific k -partition in \mathcal{P}_k which represents the set of all possible k -partitions.

Let $M := W - dd^T$ denotes the modularity matrix of $G(V, W)$. If we would like to get clusters with similar volumes then we have to add some penalty to Equation (6) hence we get the normalized Newman-Girvan cut.

$$\max_{P_k \in \mathcal{P}_k} \sum_{a=1}^k \frac{1}{\text{Vol}(V_a)} \sum_{i,j \in V_a} (W_{i,j} - d_i d_j) \quad (7)$$

where $\text{Vol}(V_a) = \sum_{u \in V_a} d_u$.

Let us define the so called normalized modularity matrix;

$$M_D := D^{-1/2} M D^{-1/2} \quad (8)$$

If we would like to cluster a weighted graph $G(V, W)$ then eigenvectors of its modularity (M) and normalized modularity matrices (M_D) can be used. Modularity and normalized modularity matrices are symmetric, and 0 is always in the spectrum of M_D .

$$M_D = \sum_{i=1}^N \lambda_i u_i = \sum_{i=1}^{N-1} \lambda_i u_i$$

where $1 > \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N \geq -1$ denote the eigenvalues of M_D .

If we would like to maximize Equation (7) we can use the k -means clustering algorithm on the optimal k -dimensional representation of vertices,

$$\left(D^{-\frac{1}{2}}u_1, \dots, D^{-\frac{1}{2}}u_k \right)^T.$$

where u_1, \dots, u_k denote the corresponding eigenvalues of $|\lambda_1(M_D)| \geq \dots \geq |\lambda_k(M_D)|$. Moreover, if the normalized modularity matrix has large positive eigenvalues, then the graph has well separated clusters, otherwise clusters are strongly connected.

Another natural approach is to minimize the normalized cut (Luxburg 2007).

$$\min_{P_k \in \mathcal{P}_k} \sum_{a=1}^{k-1} \sum_{b=a+1}^k \left(\frac{1}{\text{vol}(V_a)} + \frac{1}{\text{vol}(V_b)} \right) W_{i,j} \quad (9)$$

The optimisation problem is similar to Equation (7). Instead of the normalized-modularity matrix the normalized Laplace matrix gives the solution (Shi and Malik 2000).

$$L_D := D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} \quad (10)$$

This technique works when clusters are well separated otherwise normalized modularity gives better figures.

Algorithm

In empirical analysis, the following steps are the backbone of the calculation (Filippone et al. 2007).

1. Constructing similarity matrix (W)
2. Calculating normalized modularity matrix (M_D)
3. Based on the spectral gap, determine the number of clusters and optimal k -dimensional representation
4. Apply k -means clustering

Assessment of clustering methods

Relevance of different clustering techniques can be tested in multiple ways. The most common metrics follows a regression based logic. In this framework we suppose that variance has two components, the within and the between cluster components. Therefore, the explanatory power of given clusters can be described as

$$\frac{\sum_{j=1}^k \sum_{i=1}^{N_i} (X_{i,j} - \bar{X})^2 - \sum_{i=1}^k \sum_{j=1}^{N_i} (X_{i,j} - \bar{X}_i)^2}{\sum_{i,j=1}^{N_i N_j} (X_{i,j} - \bar{X})^2} \quad (11)$$

where k represents the number of clusters, N_i shows the size of clusters and \bar{X} , \bar{X}_i stands for the total and

clusterwise average (Zhao 2015). The formula penalizes dispersions within clusters, hence dense clusters would give number close to 1. Moreover, calculating the ratios with different number of clusters highlights the optimal number of clusters as well.

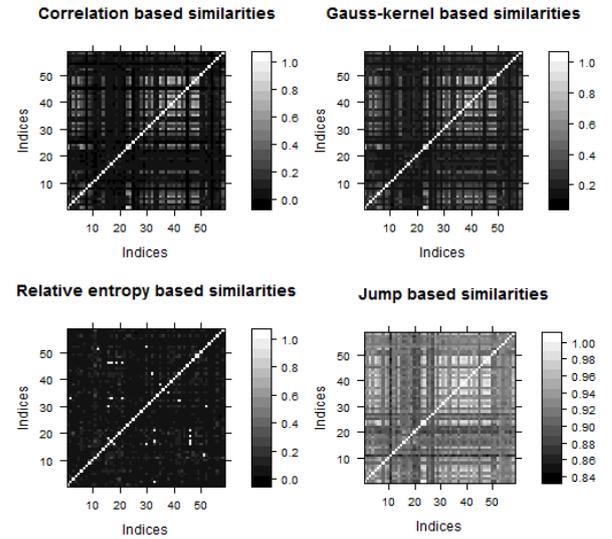
EMPIRICAL RESULTS

Current study presents a broad analysis of the equity index network structure. Logarithmic returns of 59 stock indices are clustered in different ways. The investigation reveals stock indices are homogenously connected and large price movements have limited effect on the network structure.

Similarity metrics

Defining similarity is a key aspect in clustering. In general it is hardly possible to find an optimal kernel, but different approaches can be tested and compared on specific data sets.

This study analysis correlation, jump, entropy and Gaussian based similarity kernels. Calculating the similarity matrices we expect strongly connected indices have coefficients close to one, whereas loosely connected close to zero. Level plots (Figure 1) give a feeling on the network structure which seems to be homogeneous; thus, clusters could not be well separated.



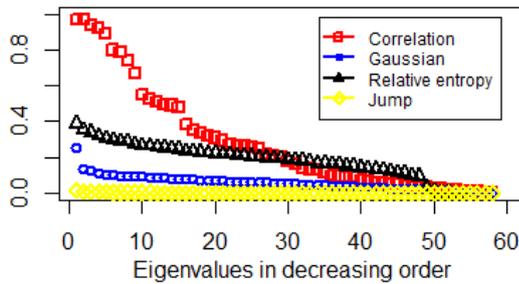
Figures 1: Levelplots of daily similarity matrices

Different similarity measures imply similar patterns which are in line with our *a priori intuition*.

However, the spectrum of normalized Laplace and normalized modularity matrices help us to find the most adequate kernel function, because the wider the spectral gap the better the clustering property. This means, we have to find similarity metrics which implies large gaps in the spectrum of normalized Laplacian and modularity matrix.

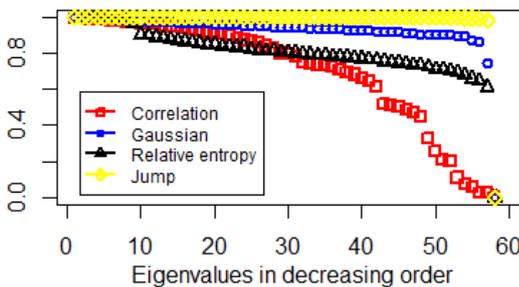
Empirical evidences (Figure 2. and Figure 3.) show relative entropy and Gaussian-kernel also can be used to cluster the stock index network, while correlation and jump based similarities are not promising.

Correlation based similarity approach implies roughly uniform eigenvalue density on $[0,1]$. This means, a lot of gaps appear in the spectrum, hence we could not say anything about the optimal number of clusters. Moreover, lower dimensional representations will not contain all the information, because of some of the large eigenvalues are not considered. These hurdles highlight the problems of squared correlation similarity matrices. Counting at least two standard deviation jumps results small number of eigenvalues with large multiplicity. Therefore, lower dimension representation can not be used to cluster the data points. Accordingly, jumps are random that do not say much about the network structure.



Figures 2: Eigenvalues of normalized modularity matrix in decreasing order

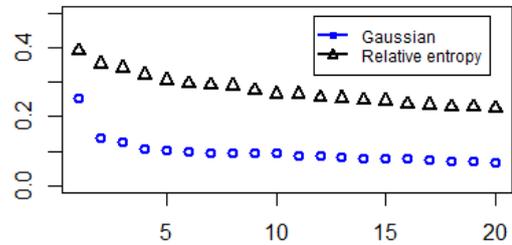
Gaussian and relative entropy based similarity matrices imply auspicious figures. Especially in normalized modularity case, we get large well separated eigenvalues, which are necessary to transform the data into a lower dimensional space.



Figures 3: Eigenvalues of normalized Laplacian matrix in decreasing order

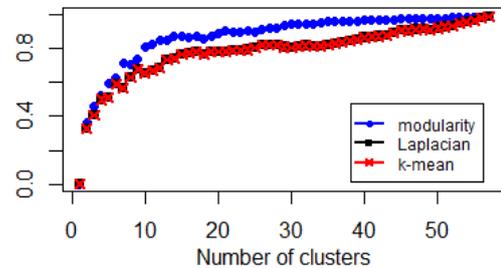
Notice that, these results are in line with Figure 1. Because, normalized Laplacian minimize the normalized cut (Equation (9)), which is small if and only if clusters are loosely connected. Whereas, modularity approach maximize the information of clustering, hence it can be used in homogeneous network structure as well.

Investigating the spectrums, especially the positions of spectral gaps, gives some guidances on the optimal number of clusters. Considering the previous results the spectrums of Gaussian and relative entropy based normalized modularity matrices are suitable. Figure 4. shows indices could be put into 2, 3 or 5 clusters.



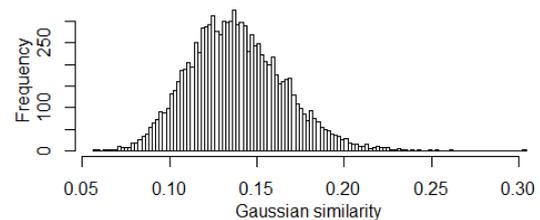
Figures 4: Large eigenvalues of Gaussian and relative entropy based normalized modularity matrices

We apply the elbow method to identify the optimal number of clusters. This approach is rather computation intensive, because of the percentage of variance explained as a function of clusters has to be estimated (Eq. 11); thus, the whole process has to be repeated many times. However; in our case we have 59 stock indices, hence the elbow method can be used as well. Figure 5. and 7. give evidences for using 2, 3, 4 or 5 clusters.



Figures 5: Explained percentage variance of Gaussian-kernel based clusters of representations

Analysing Gaussian similarity kernel shows that if we randomly generate data, then we would get similarities smaller than 0.25.



Figures 6: Histogram of 10,000 Gaussian similarities which are generated from i.i.d. 250 Dim. standard normal samples

Luxembourg	1	1	2
Canada	1	2	2
Mexico	1	2	2
Chile	1	2	2
Argentina	1	2	2
Hungary	1	1	2
Morocco	2	1	2
S&P 500	1	2	2
MSCI World	1	2	2
Czech Rep	1	1	2
Togo	2	1	2
Spain	1	1	2
Norway	1	1	2
France	1	1	2
South Africa	1	3	2
Euro Stocks	1	1	2
Sweden	1	1	2
UK	1	1	2
Netherlands	1	1	2
Finland	1	1	2
Polnad	1	3	2
Germnay	1	1	2
Belgium	1	1	2
Italy	1	1	2
Brazil	1	2	2
Colombia	1	3	2
Bangladesh	2	1	3
Costa Rica	2	1	4
Zambia	2	1	4
Malawi	2	1	4
Venezuela	2	1	4
South Korea	2	3	5
Hong Kong	2	3	5
Thailand	2	3	5
China	2	3	5
Kenya	2	3	5
India	2	3	5
Namibia	2	3	5
Turkey	2	3	5
Indonesia	2	3	5
Malaysia	2	3	5
Russia	2	3	5
Australia	2	3	5
Taiwan	2	3	5

Japan	2	3	5
Ukraine	2	3	5
Bulgaria	2	1	5
Romania	2	3	5

Notes: This table contains the list of indices and clustering results for 2, 3, and 5 clusters.

Conclusion

Spectral clustering techniques can be used to discover the equity index structure. On the one hand, clusters help us to overcome the hardship of heterogeneity. We have seen, considerable part of the total variance can be explained by them.

On the other hand various similarity approaches have unveiled tail events have little effect on the dense network structure. It has also turned out that Gaussian-kernel based clusters are in line with qualitative categorizations. In addition, clusterwise linear regressions give significant results.

All of these imply non linear effects can be eliminated by spectral clustering, thus regular mean-variance representation gives clusterwise reliable figures.

ACKNOWLEDGEMENTS

Mihály Ormos acknowledges the support by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. László Nagy acknowledges the support by the Pallas Athéné Domus Scientiae Foundation. This research is partially supported by Pallas Athéné Domus Scientiae Foundation.

REFERENCES

- B. Engelmann, E. Hayden and D. Tasche. 2003. "Measuring the Discriminative Power of Rating Systems" *Banking and Financial Supervision* No 01/2003
- G. Leibon, S. D. Pauls, D Rockmore and R. Savell. 2008. "Topological Structures in the Equities Market Network" *PNAS*, Vol. 105, 20589-20594.
- J. Shi and J. Malik. 2000. "Normalized cuts and image segmentation" *Pattern Analysis and Machine Intelligence* IEEE, Transactions, N.J., 888-905.
- M. Bolla. 2011. "Penalized version of Newman-Girvan modularity and their relation to normalized cuts and k-means clustering" *Physical review* Vol. 84.
- M. Bolla. 2013. "Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables" *Wiley*
- M. Filippone, F Camastra, F. Masulli and S. Rovetta. 2007. "A survey of kernel and spectral methods for clustering" *Pattern recognition* Vol. 41, 176-190.
- M. Ormos and D. Zibriczky. 2014. "Entropy-Based Financial Asset Pricing" *PLoS ONE* 9(12): e115742.
- R. Yalamova. 2009. "Correlations in Financial Time Series during Extreme Events" Spectral Clustering and Partition Decoupling Method" *Proc. of World Congress on Eng.*
- U. von Luxburg. 2007. "Tutorial on Spectral Clustering" *Statistics and Computing* Vol. 17, 395-416.
- Y. Zhao. 2015. "R and Data Mining: Examples and Case Study" *Elsevir Inc.*

THE USE OF CLUSTER ANALYSIS FOR DEMOGRAPHIC POLICY DEVELOPMENT: EVIDENCE FROM RUSSIA

Oksana Shubat
Ural Federal University
620002, Ekaterinburg, Russia
Email: o.m.shubat@urfu.ru

Abilova Makhabat
Magnitogorsk State Technical University
455000, Magnitogorsk, Russia
Email: abilova.mahabat@yandex.ru

Anna Bagirova
Ural Federal University
620002, Ekaterinburg, Russia
Email: a.p.bagirova@urfu.ru

Anton Ivlev
Magnitogorsk State Technical University
455000, Magnitogorsk, Russia
Email: ivlevanton@bk.ru

KEYWORDS

Cluster analysis, Demographic policy, Fertility, Birth rate, Russian regions

ABSTRACT

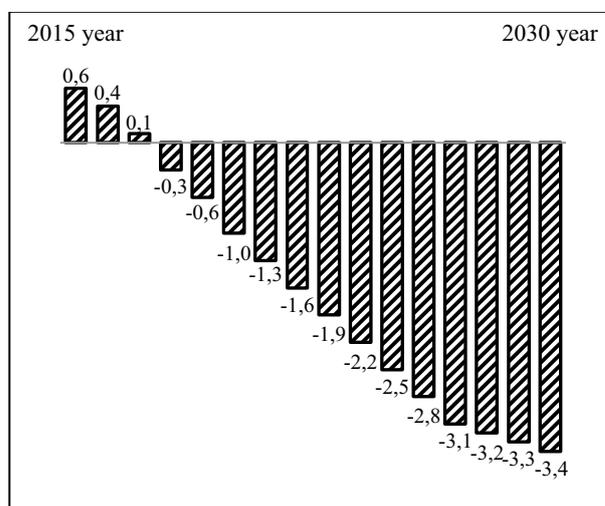
Russia has been experiencing a demographic crisis since the 1990s. The most obvious manifestations include an excess of mortality over fertility rates, population decline and an ageing population. The last 20 years have seen considerable activity to come up with new demographic policy measures to mitigate these adverse trends, with single solutions developed for all regions in Russia. This paper presents the results of a study where cluster analysis was applied to enable the identification of groups of regions with significant differences in the dynamics of socio-demographic indicators. We used hierarchical cluster analysis to classify and group Russian regions on the basis of social and economic development indices for 2002 and 2008. The validity of the profiling was confirmed using parametric and non-parametric tests. The analysis identified three clusters of Russian regions. These clusters have significant differences in socio-demographic indicators and the associated dynamics. The results of our analysis identified 'growth points' for each cluster: the fertility correlates that should be factored into the development of effective demographic policy measures.

INTRODUCTION

Since the early 1990s, Russia has been grappling with a demographic crisis. It is most evident through a mortality rate that exceeds the birth rate, population decline and an ageing population. According to forecasts by official Russian statistics body Rosstat, between 2015 and 2030, the rate of natural increase in Russia will fall from 0.6‰ to -3.4‰ (Figure 1).

There are several key reasons for low fertility in Russia. These include: a prevailing social norm of having few children, a lack of stability guarantees, structural factor influences (a declining proportion of women of a reproductive age), and an extended period

of time when the authorities paid no attention to this problem (Klupt 2008).



Figures 1: Rate of natural increase in Russia (‰): medium variant of projection by official Russian statistics (Births, deaths and natural increase 2014)

The end of the 20th century and the start of the 21st century saw an uptick in the development of new demographic policy measures aimed at mitigating these negative trends. However a single set of measures aimed at boosting fertility was proposed for different Russian regions. The most innovative was the introduction of the so-called 'maternity capital' – a one-off lump-sum amount paid for the birth of the second child. When it was introduced in 2007, this payment amounted to \$9,755; in today's terms (to account for the change in the rouble/dollar exchange rate), this is equal to \$7,215. Under Russian law, this money can be spent on improving living conditions, on the child's education or on the mother's retirement savings. Other noteworthy fertility incentives include 140-day fully paid maternity leave; part-paid maternity leave until the child is 18 months old; and unpaid maternity leave until the child turns 3. According to a number of studies, the introduction of lump sum maternity payments had an

impact on Russian birth rates, most of all in regions with a poor standard of living. In many parts of Russia, this money was enough to buy a flat, whereas in Moscow it would not get much further than a few square feet of living space.

Notably, there are important historical differences between the various parts of Russia as regards socio-economic parameters and overall standards of living. Thus the 2013 birth rate in Russia was 13.2‰, but the recorded minimum was 8.8‰ in Leningrad Region and 26.1‰ in Tyva Republic. The average gross regional product for 2013 was \$11,930 per capita, with a minimum of \$2,804 in the Republic of Chechnya and a maximum of \$45,075 in Tyumen Region. Clearly, the use of a single set of measures to address demographic problems in different parts of Russia cannot be effective. The applied measures should be tailored to match the needs of groups of regions that are in similar demographic situations and have common demographic dynamics. We believe that the statistical methodology of cluster analysis could be an appropriate instrument for doing this.

Cluster analysis is hardly used today in the development of differentiated demographic policies. Instead, there is a preference for purportedly ‘one size fits all’ mechanisms to boost fertility, despite the fact that such measures do not exist. Every country and every region should come up with a set of measures that reflect the unique combination of economic, political, socio-cultural, psychological and religious specifics of the local population. Unfortunately, the particular nature of these factors and their possible effects on the impact of demographic policy are not taken into consideration in Russia today. We believe that this could relate to a number of factors: 1) insufficient attention from the authorities to the results of demographic research in Russia; 2) variability between Russian regions, considerable differentiation in their economic and socio-cultural statuses; 3) the absence of the requisite analytical skills among officials implementing demographic policies around Russia.

Notably, in a broader context, cluster analysis is often used for the segmentation of regions. For example, O. Simpach and J. Langhamrova used it to study the impact of ICT growth on households and municipalities in regions of the Czech Republic (Simpach and Langhamrova 2014). A. Repkin identified clusters of Asian countries on the basis of economic indicators (Repkin 2012); F. Kronthaler identified clusters of German regions on the basis of economic capability (Kronthaler 2005). O. Simpach applied cluster analysis to demographic development indicators to segment municipalities (Simpach 2013). The results of cluster analysis are used today to develop “cluster strategies”, which “have become a popular economic development approach among state and local policymakers and economic development practitioners” (Cortright 2006).

Our research aims to identify groups of Russian regions that share common demographic characteristics and dynamics. They would thus benefit from the

implementation of similar demographic policy measures that are nevertheless tailored to the needs of each individual regional cluster.

DATA AND METHODS

1. We performed a hierarchical clustering procedure to identify groups of regions characterised by similar fertility trends. In carrying out this analysis, we undertook activities typical for this type of statistical work: the selection and transformation of input variables; the selection of distance measures and linkage rules; the selection of the method of clustering; the selection of the number of clusters; the profiling of clusters and interpretation of the attained results.

2. We used hierarchical cluster analysis in our research. We used squared Euclidean distance as the distance measure and Ward’s method to gauge distance between clusters. The decision on the number of identified groups of regions was taken on the basis of:

- ✓ Graphical representation of the clustering process (we examined a dendrogram);
- ✓ A coefficient showing the distance between the linked clusters;
- ✓ An evaluation of the between-group and within-group variability;
- ✓ Cluster size (we tracked the number of regions that form a single cluster to ensure that each group contained a sufficient number of regions).

3. To assess the stability and validity of the cluster solution, we performed several iterations of the clustering procedure, using different measures of distance between objects. Moreover, we applied partitioning methods of clustering (k-means procedure). For most Russian regions, their allocation into homogenous groups coincided. Some differences in cluster composition did not skew the profile of each identified group of regions. The shared characteristics and relationships identified in the course of the analysis did not change when different distance measures were used.

4. We used both the clustering variables and other variables that describe the socio-economic development in a region to interpret the clusters themselves. We examined cluster centroids, calculated the clustering variables’ average values for all objects in a certain group of regions and executed tests of significance in difference between two means. We applied one-way analysis of variance (ANOVA) to evaluate differences between the means and Levene’s test to assess the homogeneity of variances. To test the assumption of normality, we used the Shapiro-Wilks test.

5. To carry out cluster analysis, we selected five variables that provide some indication of the demographic situation in a region. The first four variables are well-known demographic indicators, presented in official Russian statistics by region:

- 1) Birth rate;
- 2) Perinatal mortality rate;
- 3) Infant mortality rate;
- 4) Under-five mortality rate.

As an additional clustering variable, we introduced a calculated value: pregnancy rate. This is the number of completed or terminated pregnancies per 1,000 women aged 15-49. We believe that this variable enables us to assess the relative level of reproductive activity within a particular community. Thus we believe that the coefficient of pregnancy rate, along with other demographic indicators could be used as an informational basis for developing effective demographic policy.

Another variable that also characterises the demographic situation is the number of abortions per 100 births. The Russian State Statistics Service provides this data both at a country and a regional level, enabling us to use it in our analysis. However we did not use this variable for our clustering, as it clearly correlates to other variables in our study, including pregnancy rate. At the same time, we used the number of abortions per 100 births for subsequent interpretation of the identified clusters.

6. Throughout the clustering, we did not use the variables themselves but rather their indices: values that describe changes in the demographic variables. We believe that studying the dynamics of demographic processes, rather than their static values, enables a more profound understanding of the demographic situation in a region. Similarly, the identification of regional groups that have similar trends in the development of demographic processes is more justified than basing the groups on attained (transpired) demographic indicators. In our view, developing state fertility support measures that account for the dynamics of population replacement enables greater effectiveness. We also note that since all of the input variables we drew on in the clustering were indices, we did not need to standardize the variables beforehand.

7. To carry out the hierarchical clustering procedure, we used variables that characterise the demographic situation in Russian regions between 2002 and 2008. This was not an accidental choice: active state support and the incentivizing of fertility started in Russia in 2000. In response to dramatic depopulation challenges, the government developed a new Demographic Policy Concept and began its implementation in late 2000-early 2001. Thus we felt it would be quite fitting to analyse the possible results of this new demographic policy from 2002. The subsequent years – up to and including 2008 – can be described as a period of relative macroeconomic stability. However the financial crisis that hit most intensely in early 2009 had a negative impact on reproductive indicators in Russia. We believe that these “fertility shocks” should be excluded from the sample in our analysis and studied as a separate phenomenon. Thus our analysis focused on data for 2002-2008.

8. For the purposes of the cluster analysis, our study sample included all Russian regions that had complete data for all input variables. We note that the time between 2003 and 2008 saw a great deal of political and economic activity around the administrative

consolidation of neighbouring territorial units with tight economic links. As a result of this process, the number of Russian constituent federal parts went from 89 to 83. For our hierarchical clustering procedure, we only selected regions that were not affected by the consolidation processes. Thus 78 Russian regions were included in the cluster analysis.

RESULTS

In performing the cluster analysis, we settled on a three-cluster solution. The first cluster included 37 Russian regions; there were 14 regions in the second and 27 in the third (Figure 2). The evaluation of the cluster centroid confirmed the appropriateness of these three groups: the mean and median values of the cluster variables differed significantly between the identified clusters and, in most cases, when compared to the nationwide levels (table 1). Levene's test confirmed the validity of the one-way ANOVA (table 2). The Shapiro-Wilks test confirmed that each of the levels of the variables is normally distributed (table 3). Results of the one-way ANOVA are presented in Table 4.

Table 1: Statistical characteristics for the cluster variables in 2008

		Cluster 1	Cluster 2	Cluster 3
Birth rate	Mean	11.9	12.3	13.3
	Median	11.5	11.8	12.5
Perinatal mortality rate	Mean	7.6	12.6	8.9
	Median	7.1	12.0	8.5
Infant mortality rate	Mean	8.5	9.2	8.5
	Median	8.1	8.9	8.2
Under-five mortality rate	Mean	8.3	12.5	8.4
	Median	8.1	11.7	7.9
Pregnancy rate	Mean	83.2	87.5	75.6
	Median	79.9	82.3	75.0

Table 2: Test of homogeneity of variances

Variables	Levene Statistic	df1	df2	Sig.
Birth rate	0.770	2	75	0.467
Perinatal mortality rate	0.045	2	75	0.956
Infant mortality rate	2.301	2	75	0.107
Under-five mortality rate	2.049	2	75	0.136
Pregnancy rate	1.367	2	75	0.261

Table 3: Test of normality

Variables	Clusters	Shapiro-Wilk		
		Statistic	df	Sig.
Birth rate	1	0.966	37	0.319
	2	0.958	27	0.331
	3	0.919	14	0.213
Perinatal mortality rate	1	0.970	37	0.413
	2	0.964	27	0.457
	3	0.905	14	0.135
Infant mortality rate	1	0.958	37	0.170
	2	0.950	27	0.220
	3	0.902	14	0.120
Under-five mortality rate	1	0.943	37	0.059
	2	0.975	27	0.742
	3	0.879	14	0.056
Pregnancy rate	1	0.976	37	0.581
	2	0.967	27	0.520
	3	0.932	14	0.328

Table 4: ANOVA

	Between Groups	Within Groups	Total
Birth rate			
Sum of Squares	31.399	105.228	136.627
df	2	75	77
Mean Square	15.7	1.403	
F	11.19		
Sig.	0.000		
Perinatal mortality rate			
Sum of Squares	275.274	46.521	321.795
df	2	75	77
Mean Square	137.637	0.62	
F	221.893		
Sig.	0.000		
Infant mortality rate			
Sum of Squares	61.616	14.029	75.645
df	2	75	77
Mean Square	30.808	0.187	
F	164.699		
Sig.	0.000		
Under-five mortality rate			
Sum of Squares	160.622	33.262	193.884
df	2	75	77
Mean Square	80.311	0.443	
F	181.09		
Sig.	0.000		
Pregnancy rate			
Sum of Squares	2870.706	8268.406	11139.112
df	2	75	77
Mean Square	1435.353	110.245	
F	13.02		
Sig.	0.000		

The results of the clustering and the subsequent interpretation showed that the clusters differ in:

- 1) The values of the evaluated demographic variables;
- 2) The nature of the dynamics of the evaluated demographic variables;
- 3) The level of economic development.

Cluster 1 – “Low fertility amid low economic activity” This is the largest cluster comprising 37 regions. It covers approximately 44% of the territory of Russia and accounts for around half of the country’s population. Its most outstanding demographic characteristics include:

- for the entire period of the study, the fertility rate in the regions of this cluster was below that of the other two clusters and lower than the national average;
- the pregnancy rate for the entire period was below the national average;
- the number of abortions per 100 births was high and exceeded the national average.

Moreover, regions in the first cluster also shared the following characteristics of the index dynamics:

- a growth in the fertility rate (in 2002 this was 9.7‰, rising to 11.9‰ in 2008);
- a decline in the number of pregnancies (88.7% in 2002 to 84.2% in 2008);
- a decline in the number of abortions (143 abortions per 100 births in 2002; 89 abortions/100 births in 2008);
- a certain lag behind national dynamics (for example, the total fertility rate in regions of this cluster for the studied period grew by 22%, while the increase across all of Russia was 25%; the decline in abortions per 100 births for this cluster fell by 38% compared to 42% in Russia overall).

Interpretation of this cluster through variables that reflect the level of economic development showed that this cluster includes a dominant share of people employed in the economy, and also produces the greatest volume of manufacturing. This cluster is also characterised by the largest share of commissioned residential buildings (almost half of the total quantity for Russia). However, despite the significant size of the first cluster, a number of economic indicators were not particularly high. These included the value of fixed assets, volume of natural resource production, import and export turnover, contribution to GNP and gross national income. It can be said that regions of the first cluster are not economically active as regards a number of key economic factors.

Cluster 2 – “Cautious” fertility amid high economic activity”

This cluster comprises 14 Russian regions, which cover a quarter of the total area of the country. The total population accounts for one-fifth of the total population. The key characteristics of this cluster include:

- the highest pregnancy rate across the clusters;
- the lowest number of abortions per 100 births across the clusters;

- the highest perinatal and infant mortality rates across the clusters;
- a higher level of under-five mortality compared to other clusters.

Other characteristics of the demographic situation in regions of this cluster were not generally outstanding and were quite similar to pan-Russian trends. Thus the total fertility rate for this cluster for the entire study period virtually matched the national rate; there was an overall tendency towards a decrease in mortality rates, in line with the overall trend for Russia.

Interpretation of the cluster through variables related to economic development showed that this cluster did not have a significant share of people employed in the economy (around one-fifth of the total active population). At the same time, the cluster leads on contribution to GNP (the total gross regional product for this cluster was approximately 42% of GNP). Regions of this cluster also account for a majority stake in natural resource production industries (around 60% of total Russian volumes), substantial fixed asset value, and export and import turnover. Moreover, the share of investments in fixed assets in this group of regions was almost one-third of the figure for Russia. Thus as far as key economic indicators go, this cluster can be seen as relatively successful and extremely vigorous as regards economic activity. The economy in this cluster is largely geared towards innovative development.

Cluster 3 – “High fertility amid economic passivity”

This cluster includes 27 Russian regions and comprises around one-third of the Russian population. Its key characteristics include:

- a fertility rate that exceeded that of other clusters and the national average for the duration of the study;
- a pregnancy rate that was below the other clusters and below the national average for the duration of the study.

Moreover, regions in this group were also characterised by the following demographic dynamics:

- the highest growth in fertility rates among the clusters;
- an insignificant decline in the pregnancy rate;
- a significant decrease in the number of abortions: from 111 abortions per 100 births in 2002 to 67 abortions per 100 births in 2008.

Other demographic characteristics in this cluster were not particularly remarkable and were very close to overarching trends for Russia.

Interpretation of this cluster through economic development variables showed that this group accounts for approximately one-third of people employed in the economy. Most of the analysed economic indicators had the lowest values in this cluster. Thus the share of the gross regional product was approximately one-fifth of GNP. Natural resource production volumes, fixed asset value, retail trade turnover, and import and export volumes were similarly low compared to the other

clusters. At the same time, as regards a number of other economic factors, the third cluster was an undisputed leader. This concerns agriculture (43% of the total volume for Russia), plant products (45%) and animal products (41%). On the whole, the third cluster is characterised by a certain economic passivity and even depression against a backdrop of considerable agricultural activity.

DISCUSSION

The results of our cluster analysis show that regional specifics should be factored into fertility research. This is hardly surprising, given that there is no single theory for stimulating fertility in the modern world. On the contrary, there are a great many diverse approaches for mitigating low fertility rates being implemented around the world, with varying degrees of success. Moreover, the replication of one country’s (or one region’s) successful strategy in another location is unlikely to be effective. The specific nature of the demographic, economic and political trends of a country or region, the nuances of local social and moral norms and processes should be taken into account in the development of fertility incentive frameworks.

However in Russia, maternity capital payments for the birth of the second child remain the key measure for stimulating fertility. We are inclined to criticise this measure for the following reasons. Firstly, a two-child family should not be the priority for Russia at this point in time. Between January and May 2015, the natural decrease was several times higher than in previous years. The economic crisis is likely to have exacerbated this adverse dynamic. Undoubtedly, it will continue to have a delayed negative impact in the future as well. The situation is made worse by changing age structures. According to Rosstat, by 2025, the number of women of childbearing age in Russia will fall by 7.2%, and by 2030 – by 10.3%. The share of women of fertile age in the population will decrease from the current 21.6% today to 19.2% in 2030. In this context, state policy should focus on increasing the number of three and four-child families, rather than two-child ones. Secondly, rewarding the quantity of children without regard for their “quality” will undoubtedly lead to deterioration in the quality of the population: a less healthy nation with lower levels of educational and cultural standing. Thirdly, financial incentives for the mere act of childbirth without accounting for subsequent activities around children’s upbringing and socialisation suggests that the state is simply interested in having more citizens, irrespective of their development and that state policy does not view parenting as a multi-faceted process that is difficult to carry out well.

The results of our analysis have allowed us to suggest measures that would be effective in boosting fertility and parenting in each cluster. For instance, the first cluster would benefit from better socio-economic conditions for childbirth and parenting at a regional level. In particular, this should focus on improving

organisational and medical facilities for families, like those envisaged in the Russian state Demographic Policy Concept through 2025 (Demographic Policy Concept 2007):

1) Improved accessibility and quality of free medical services for women for pregnancy and childbirth, and for newborn children;

2) Improved material and human resourcing for motherhood and childhood services;

3) The development of high-tech medical assistance for women during pregnancy and childbirth, and for newborn children;

4) The introduction of integrated measures for further reduction in the number of abortions.

Undoubtedly, these measures are important for all territories, but the results of our analysis showed that they are of the highest priority for regions in the first cluster.

Let us furnish another example of the use of our analysis for regions of the second cluster. To stimulate fertility here, we propose introducing regional tax concessions for parents and creating an integrated set of special legislative conditions to support parenting at the regional and municipal levels. The innovation-focused socio-economic development of these regions will drive the evolution of skills like logical and abstract thinking, an analytical mind-set, the ability to filter and process large volumes of information and an aptitude for creativity. This will lead to more complex, differentiated and diverse work, in turn prompting changes in the structure and quality of workers and expectations about their professional competencies and traits like accountability, self-motivation, creativity and overall effort at work. We believe that these are the very regions that would benefit from a system of training for parents aimed at developing special competencies: the skills required to operate, build relationships and work in an innovation-centric economic. Parents should be taught how to help their children be creative, develop analytical skills, cope with multitasking and so on: that is, to establish the very personal qualities that will be sought after in regions of the second cluster in the near future.

Thus the results of our cluster analysis across Russian regions helped us to identify high-priority instruments for stimulating fertility and parenting, which are shaped to the needs of each type of region. Applying these as part of regional strategies enables:

1) Creating a tailored approach to mitigate mismatches between actual and forecast population numbers and composition in a particular region;

2) Developing and testing the legislative, economic and methodological bases for using a set of instruments to stimulate fertility with a defined set of regions;

3) Targeting the demographic situation in different regions with account of the identified specifics;

4) Creating conditions for the growth of birth rates in different types of regions through the use of different types of incentive measures

CONCLUSIONS

Our research showed that countries composed of many constituent parts with a high degree of variability as regards socio-economic and demographic development and the nature of the dynamics of key socio-economic indicators require demographic policies that are differentiated by region type. The regional groups should be identified on the basis of static key indicators as well as the nature of the dynamics of key indicators. Cluster analysis is an effective analytical instrument for carrying out regional segmentation in order to develop tailored demographic policy measures. Indeed, on the basis of analysis that highlighted key problems in each cluster of regions, we proposed a redistribution of resources allocated to implementing demographic policy measures. We tried to identify the measures that would not only allow regions to improve fertility in the future, but also raise the qualitative indicators of the population.

We expect to extend our research by further differentiating the regions within each cluster. Thus, there is scope to identify sub-clusters with indicators above and below median values – both for input variables and other variables related to the assessment of the socio-economic situation in the regions. This will allow ascertaining the most challenging Russian territories in each segment, which require priority measures for stimulating fertility and supporting parenting. These most disadvantaged of regions could become pilot projects for focusing state policy in the short term.

ACKNOWLEDGMENTS

The article is processed as one of the outputs of the research project “Integration of the parental labor results in Russian pension system“, supported by the Russian Foundation for Humanity, project no. 16-32-00020. This project is co-financed by Ural Federal University (Act 211 Government of the Russian Federation, contract № 02.A03.21.0006).

REFERENCES

- Births, deaths and natural increase. 2014. Demographic Projections by the Federal State Statistics Service. Rosstat, Moscow.
- Cortright, J. 2006. *Making Sense of Clusters: Regional Competitiveness and Economic Development*. The Brookings Institution, Washington, D.C.
- Demographic Policy Concept of the Russian Federation until 2025. 2007. Approved with Presidential Decree from 9 October 2007 No. 1351.
- Klupt, M.A. 2008. *Demographics of regions of the world*. Piter, Saint Petersburg.
- Kronthaler, F. 2005. “Economic Capability of East German Regions: Results of a Cluster Analysis”. *Regional Studies*, Vol 39, Issue 6, 739-750.
- Repkine, A. 2012. “How Similar Are the East Asian Economies? A Cluster Analysis Perspective on

Economic Cooperation in the Region”. *Journal of International and Area Studies*, Vol 19, Issue 1, 27-44.

Simpach, O. 2013. “Application of Cluster Analysis on the Demographic Development of Municipalities in the Districts of Liberecky Region”. In *Conference Proceedings of the 7th International Days of Statistics and Economics* (Prague, Czech Republic, Sept. 19-21). Melandrium, 1390-1399.

Simpach, O. and Langhamrova, J. 2014. “The Impact of ICT Growth on Households and Municipalities in the Czech NUTS-3 Regions: the Application of Cluster Analysis”. *Schriftenreihe Informatik*, Vol 43: IDIMT-2014: Networking Societies - Cooperation and Conflict, 63-70.

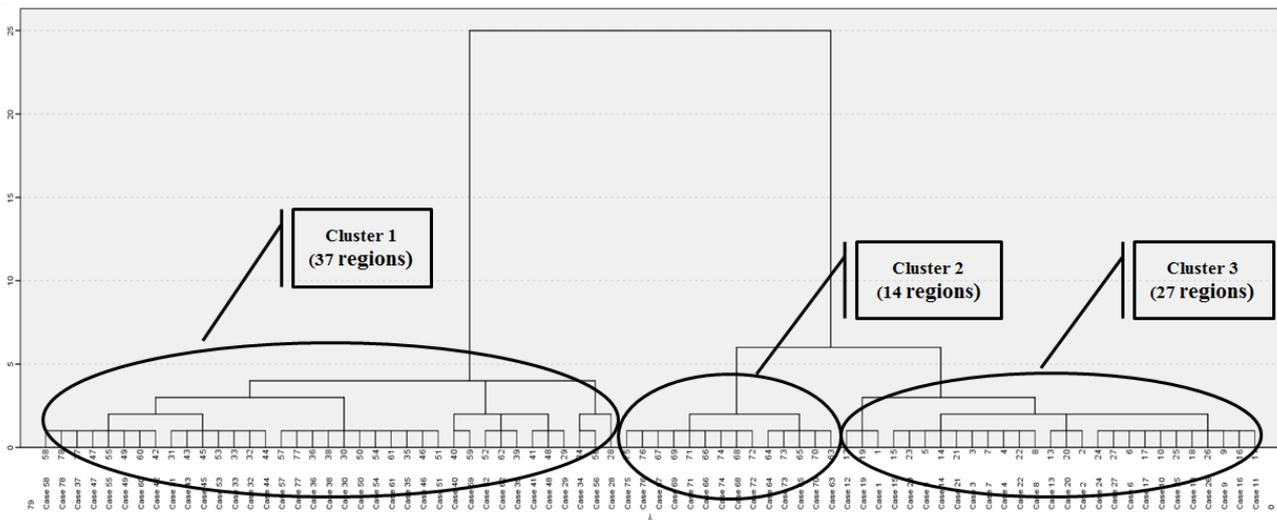
AUTHOR BIOGRAPHIES

OKSANA SHUBAT is an Associate Professor of Economics at Ural Federal University (Russia). She has received her PhD in Accounting and Statistics in 2009. Her research interests include demographic processes, demographic dynamics and its impact on human resources development and the development of human capital (especially at the household-level). Her email address is: o.m.shubat@urfu.ru and her Web-page can be found at <http://urfu.ru/ru/about/personal-pages/O.M.Shubat/>

ANNA BAGIROVA is a professor of economics and sociology at Ural Federal University (Russia). Her research interests include demographical processes and their determinants. She also explores issues of labour economics and sociology of labour. She is a doctoral supervisor and a member of International Sociological Association. Her email address is: a.p.bagirova@urfu.ru and her Web-page can be found at <http://urfu.ru/ru/about/personal-pages/a.p.bagirova/>

MAKHABAT ABILOVA is an Associate Professor of Economics at Magnitogorsk State Technical University (Russia). She has received her PhD in Economics in 2010. Her research interests include the analysis of human capital and statistical methods of demography. Her email address is: abilova.mahabat@yandex.ru

ANTON IVLEV is an Associate Professor of Economics at Magnitogorsk State Technical University (Russia). He has received his PhD in Economics in 2011. He explores issues of labour economics and financial model of education. His email address is: ivlevanton@bk.ru



Figures 2: Dendrogram using Ward Linkage

CLASSICAL AND NOVEL RISK MEASURES FOR A STOCK INDEX ON A DEVELOPED MARKET

Nagy Júlia Tímea

Business analyst, Wolters Kluwer
Financial Services,
Bulevardul 21 Decembrie 1989, Nr. 77 |
400124 Cluj | Romania Cluj-Napoca, Romania

Nagy Bálint Zsolt,

Associate professor, Babes-Bolyai University, Faculty of
Economics and Business Administration, Str. Teodor
Mihali, Nr.58-60, 400591, Cluj-Napoca, Romania,
corresponding author: nagybzsolto@yahoo.com

Juhász Jácint

Lecturer, Babes-Bolyai University, Faculty of Economics
and Business Administration, Str. Teodor Mihali, Nr.58-
60, 400591, Cluj-Napoca, Romania,

Abstract

In the present article we conduct an inquiry into several different risk measures, illustrating their advantages and disadvantages, regulatory aspects and apply them on a stock index on a developed market: the DAX index. Specifically we are talking about Value at Risk (VaR), which is now considered a classical measure, its improved version, the Expected Shortfall (ES) and the very novel Entropic Value at Risk (EvaR). The applied computation methods are historic simulation, Monte Carlo simulation and the resampling method, which are all non-parametric methods, yielding robust results. The obtained values are put into the context of the relevant literature, and pertinent conclusions are formulated, especially regarding regulatory applications.

Keywords: Value At Risk, Expected Shortfall, Entropic Value At Risk, simulation, resampling

JEL classification: C14, C15, G11

1. Literature review

In the present chapter we introduce three risk measures: VaR – Value at Risk, ES – Expected Shortfall or otherwise called CVaR – Conditional Value at Risk, and EVaR – Entropic Value at Risk. In the classification given in Albrecht (2003), these measures are all part of the category of absolute risk measures. We examine to what extent these measures can be considered coherent, at the same time we analyse how much these risk measures differ in terms of figures, and highlight which ones were the most realistic. Finally, we examine how the more novel risk measures such as ES and EVaR brought new insights and ameliorated certain aspects of classical VaR.

Because of length considerations we don't have the space to define VaR and ES here, we consider them widely known, we shall only pertain to entropic VaR.

1.1.EVaR (Entropic Value at Risk)

Definition of EVaR (based on Ahmadi-Javid, 2012): Let (Ω, \mathcal{F}, P) be a probability space with Ω a set of all simple events, \mathcal{F} a σ -algebra of subsets of Ω and P a probability measure on \mathcal{F} . Let X be a random variable and L_{M^+} be the set of all Borel measurable functions $X: \Omega \rightarrow \mathbf{R}$ whose moment-generating function $M_X(z)$ exists for all $z \geq \mathbf{0}$. The entropic value-at-risk (EVaR) of $X \in L_{M^+}$ with confidence level $(1 - \alpha)$ is defined as follows:

$$EVaR_{1-\alpha}(X) := \inf \{ z^{-1} \ln(M_X(z) / \alpha) \} \quad \forall z > 0 \quad (1)$$

In finance, the random variable $X \in L_{M^+}$, in the above equation, is used to model the *losses* of a portfolio or stock index. Consider the Chernoff inequality (Chernoff, H. (1981))

$$P(X \geq a) \leq e^{-za} M_X(z) \quad \forall z > 0 \quad (2)$$

Solving the equation $e^{-za} M_X(z) = \alpha$ for a , results in $a_X(\alpha, z) = z^{-1} \ln(M_X(z) / \alpha)$. By considering the equation (1), we see that $EVaR_{1-\alpha}(X) = \inf \{ a_X(\alpha, z) \}$, $\forall z > 0$ which shows the relationship between the EVaR and the Chernoff inequality. Here $a_X(\mathbf{1}, z)$ is the *entropic risk measure* (the term used in finance) or *exponential premium* (the term used in insurance).

In case of the normal distribution ($X \sim N(\mu, \sigma^2)$), this reduces to the closed form formula:

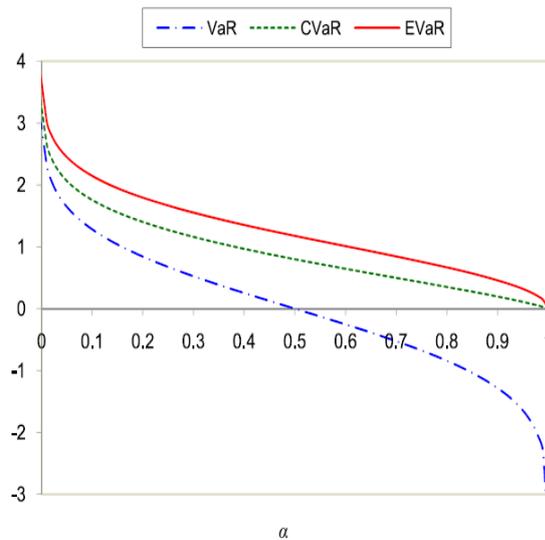
$$EVaR_{1-\alpha}(X) = \mu + \sqrt{-2 \ln \alpha} \sigma \quad (3)$$

In case of the uniform distribution, $X \sim U(a, b)$:

$$EVaR_{1-\alpha}(X) = \inf \left\{ t \ln \left(t \frac{e^{t^{-1}b} - e^{t^{-1}a}}{b - a} \right) - t \ln \alpha \right\}$$

The next chart shows the three types of value at risk in case of the standard normal distribution.

Chart 1: VaR, CVaR and EVaR in case of the standard normal distribution



Source: Ahamdi-Javid A., 2012

Comparing the three types of VaR, it can be seen that EVaR is more conservative (risk averse) than CVaR and VaR. This means that EVaR will prescribe higher capital buffers for financial and insurance institutions (this makes EVaR less attractive for certain institutions). Having a closed form formula in case of certain distributions, EVaR is more computationally tractable than CVaR (which makes it more useful for stochastic optimization problems).

The main advantage of simple VaR is that it performs a full characterization of the distribution returns, leading also to improved performance. The main disadvantage of it is that in reality the market returns are not normally distributed. The main advantage of ES is that it captures the average losses exceeding VaR, the main disadvantage being the fact that it is not easily computationally tractable.

1.2. Review of some empirical results about value at risk

We report the following interesting results in the literature regarding entropic VaR:

Rudloff et al (2008) consider the optimal selection of portfolios for utility maximizing investors under both budget and risk constraints. The risk is measured in terms of entropic risk. They find that even though the entropic risk (ER)-optimal portfolio and the pure stock portfolio coincide w.r.t. entropic risk (since the risk bound was chosen in this way), the ER-optimal portfolio clearly outperforms the stock portfolio w.r.t. expected utility. Long and Qi (2014) study the discrete optimization problem by optimizing the Entropic Value-at-Risk, and propose an efficient approximation algorithm to resolve the problem via solving a

sequence of nominal problems. Zheng and Chen (2012) propose a new coherent risk measure called iso-entropic risk measure. Comparing with several current important risk measures, it turns out that this new risk measure has advantages over the others. It is coherent, it is a smooth measure, and it is a more prudent risk measure than CVaR.

Dargiri M. N. et al (2013) compared VaR and CVaR in Malaysian sectors and based on the KLCI (Kuala Lumpur Composite Index) benchmark stock index. They used daily data covering the 2002-2012 period. They calculated VaR with the parametric, delta-normal, ARCH (Autoregressive Conditional Heteroskedasticity)-based, historic and Monte Carlo simulation methods and CVaR with delta normal and historical simulation. The study analysed eight sectors and according to the VaR calculated with all the methods listed above, the technology sector proved to be the most risky, in contrast with the construction sector (which was the most risky according to CVaR). Using a confidence level of 95% and a one day holding period, VaR reached the biggest values in case of the Monte Carlo simulation, and CVaR proved to be more conservative than VaR, having bigger values for all sectors and indices.

Čorkalo, S. (2011) used the variance-covariance method, historic and Monte Carlo simulation with bootstrapping for calculating VaR. Bootstrapping is an alternative to random number generation which relies on random sampling with replacement. The article studied 5 stocks from the Zagreb Stock Exchange and data between 2008-2010: the biggest values were obtained in the case of historic simulation, followed closely by the Monte Carlo with bootstrapping method.

Iqbal J et al (2010) used parametric (EQWMA¹, EWMA² and ARCH) and non-parametric (historic simulation and bootstrapping) methods for calculating VaR for the KSE100³ index. They used daily log-returns between June 1992 and June 2008. At the 95% confidence level, the VaR calculated with historical simulation was 2.524%, whereas the VaR obtained with bootstrapping was 2.547%. Similarly, at 99% confidence level the VaR obtained with bootstrapping was bigger than the one obtained with historic simulation.

Dutta, D. & Bhattacharya, B. (2008) used historic simulation and bootstrap historic simulation for VaR in the case of the S&P CNX Nifty Indian stock index, using data between April 2000 and march 2007 at a confidence level of 95%. The authors find a bigger VaR in case of the bootstrap method and at the same time favour this method compared to the simple historic simulation based on its small-sample behaviour.

¹ Equally Weighted Moving Average

² Exponentially Weighted Moving Average

³ Karachi Stock Index, Karachi, Pakistan

2. Data and methodology

In this study we calculate for the German DAX stock index the three risk measures that we presented (VaR, CVaR and EVaR) applying three methods: historical simulation, Monte Carlo simulation and resampling. We are aiming at comparing the values obtained and the methods.

The DAX index comprises the 30 biggest (in terms of capitalization and liquidity) German stocks on the Frankfurt Stock Exchange. The DAX, introduced in 1987, is a total return index, i.e. it reflects the dividends paid out by the listed companies, at the same time being well-diversified (ranging from vehicle to chemical industries, banking etc).

We used daily index quotations from January 2 2009 to 17 April 2014⁴, calculating their logarithmic return for 1357 days. Apart from the descriptive statistics we also tested the normality of the distribution of returns. The conducted normality tests were: Jarque-Bera, Doornik-Hansen, Shapiro-Wilk and Lilliefors. At a 95% confidence level all tests were significant (they all had low “p-levels”, thus the null hypothesis of normality can be safely rejected (this was already hinted by the skewness and kurtosis of the distribution).

2.1. The applied methodology

Historical simulation

Historical simulation, being one of the simplest methods for calculating VaR, is obtained by sorting ascendingly the returns of the historical period, and then taking the quantile for the given confidence level. On the other hand, CVaR (ES) can be calculated by sorting ascendingly the returns, and then taking the average of the returns exceeding VaR at the given confidence level. We calculated VaR in Microsoft Excel and ES in the “R” statistical software. We fed the data into the vector of returns, saved the length of the vector in variable „N”, and the 5% quantile into variable M. We performed iterative operations to sort descending the vector of returns. In another iteration (“for” cycle) we saved these values into a vector containing averages and calculated the average of the average vector. This gave us the ES at 95% confidence level.

An advantage of this method is that it doesn't necessitate the normality of the distribution of returns. A disadvantage is that it is based on historical data, hence it assumes that the past repeats itself.

Monte Carlo simulation

The method consists of the following steps: first we generate a big amount of random numbers (r_t), then we generate returns out of them, finally arriving at the index prices with the following formula:

$$P_t = P_{t-1} \exp(r)$$

The next step is to sort descending the index prices, into different bins, construct their histogram, then

calculate the middle of their intervals, which we multiply by their probability. Then we calculate the price for a given confidence level with the NORMDIST() function, then we sum up the negative values beginning from this value. This will give the value of VaR.

The advantage of this method is that it is also suitable when we want to calculate VaR for a longer period, and in this case the historical simulation would be too volatile.

Resampling (bootstrap historical simulation) method⁵

The method's main idea is that using historical data we can generate thousands of samples, reusing the historical data assuming the repetitiveness of data. We once again used the “R” statistical software: In case of VaR we saved the returns into a vector, saved its length and the number of samples into other two variables. Next we defined an iteration up until the number of samples, and determined the value of VaR for each sample, similarly to the historic simulation, i.e. the quantile at 95% confidence level, from the descending sorted returns. In the case of ES the commands were slightly modified. Here we saved the number of data in an „M” variable, and because ES is a sort of an “average of VaR's”, we had to apply a double iteration (a “for” cycle inside another) to the descending sorted returns and take the average of the vector of data for each sample.

EVaR was calculated using formula (3) from Ahmadi-Javid A. (2012) so we only needed the mean and the standard deviation of returns for each sample which we fed into the formula. It can be seen that this resampling method is a hybrid between historical and Monte Carlo simulation. Its advantage is that it doesn't assume anything about the distribution of data, and also that one can generate arbitrarily large samples which is very useful when there's very little historical data.

3. Empirical results

3.1 Historical simulation

Classical VaR was calculated in Microsoft Excel: for the available data we calculated its 95% quantile, (68 in our case), then with the aid of the SMALL() function we calculated the corresponding return and this gave us the 95% VaR. The value of CVaR (ES) was more easily calculated in the R statistical programme.

EVaR was determined with the Ahmadi-Javid A. (2012) formula employing the daily average return and standard deviation, at a 95% confidence level.

In the next table we present the risk measures calculated with historical simulation.

⁴ Source: <http://www.dax-indices.com>

⁵ Also called resampling with replacement

Table 1: Risk measures calculated with historical simulation at 95% confidence level for the DAX index between 2009.01.02 and 2014.04.17 (%)

Measures	VaR(95%)	ES(95%)	EVaR(95%)
Values	-2.29%	-3.29%	3.46%

Source: authors' calculations

The figure for classical VaR bears the following interpretation: under normal market conditions there is a 5% likelihood of a daily loss exceeding 2.29%.

ES can be interpreted the following way (in the pessimistic view): there is 5% chance of facing losses greater than VaR, i.e. in this 5% case the average loss of the index will be 3.29%.

As shown in the table, EVaR is always non-negative (but it's still regarded as a loss), and it gives the most stringent (conservative) risk measure compared to the other two, its interpretation being that there's a 95% chance that the daily loss will not exceed 3.46%.

3.2 Monte Carlo simulation

In this case we generated 10000 normally distributed random numbers and calculated the log-returns for them, using the daily, historical mean return and standard deviation. Then we arrived at the simulated index prices the following way:

$$r \log(i) = (\alpha * t) + (\sigma * t) * rand(i)$$

$$S(i) = r \log(i) * S_0$$

Where S(i) = the simulated price at moment „i“,

r = simulated return

α = mean return

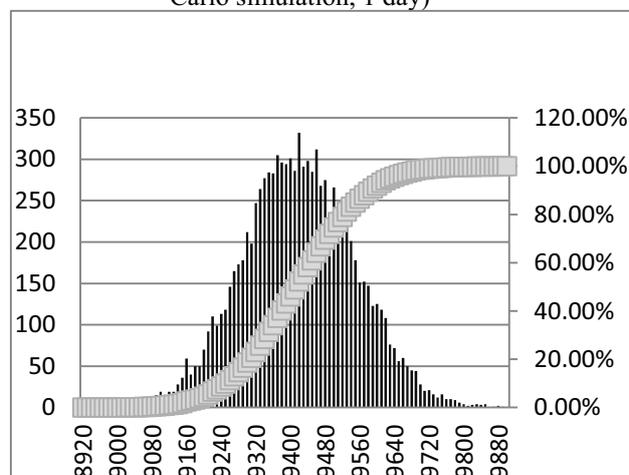
σ = standard deviation

t = daily period

i=1 → 10000 (index)

In the next step we divided the simulated index prices from the minimal (8923 euro) to the maximal (9889 euro) value and constructed the division of prices on a 10 euro scale. The next chart shows the obtained histogram and distribution function.

Chart 2.: Histogram of the simulated prices (Monte Carlo simulation, 1 day)



Runs of the script	VaR (95%)
First	-2.266%
Second	-2.267%
Third	-2.267%
Fourth	-2.266%
Fifth	-2.266%

Table 2: VaR for different runs of the resampling script Source: authors' calculations in "R"

Next, using the NORMDIST() function we calculated the return corresponding to the 95% probability which gave us the value of VaR (2.24% which means that under normal market conditions the daily loss should not be higher than 2.24% with a probability of 95%). We also calculated VaR in absolute index values. We calculated the interval mean points for the divisions and their probability. Next we filled in zeroes until the target index price (because it is unknown what happens beyond VaR), then we subtracted today's price from the interval mean, and multiplied the result by its probability. Finally beginning with the target price we summed up the negative values which gave us the VaR in absolute figures. According to this under normal market conditions the daily loss on the DAX index should not be higher than 37.33 euro with a probability of 95% (the finer the division of the intervals, the more accurate VaR will be).

CVaR (ES) was once again determined with the "R" script already presented in the case of historical simulation. The average of ES turned out to be 2.80%, which means that under normal market conditions, if the 5% probability event occurs, then the average loss will be 2.80% daily.

For EVaR we used the Ahmadi-Javid A. (2012) formula, but this time the daily average return and standard deviation were obtained from 10000 simulated returns. This gave us an EVaR of 3.42% meaning that the daily expected loss with 95% probability is 3.42%.

3.3 Resampling (Bootstrap historical simulation)

In this case we determined all three risk measures with the R statistical software, and we employed a sample size of 10000 observations.

For classical VaR we calculated its value for each sample at 95% confidence level, and the final value with resampling was given by the average of the sample VaR's. Also, each time we ran the script, we arrived at a different average VaR (see the table below, where we applied the script five times, with very small differences between the runs). We will use further the first value for comparisons between the different VaR methods.

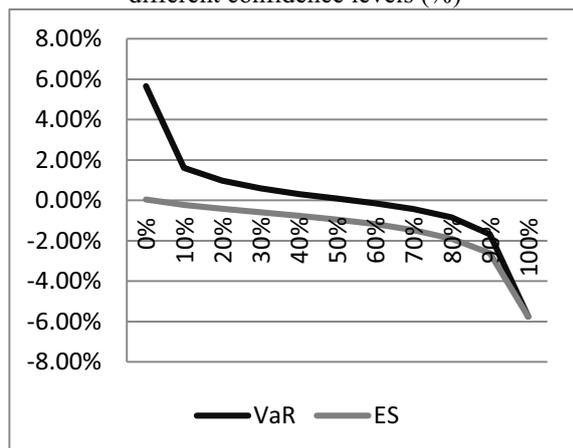
Overall, according to the resampled VaR, under normal market conditions, the average daily loss on the DAX will not be higher than 2.26% with 95% probability.

The sample size for ES was also 10000, and we saved the ES values into a vector, the average of which gave us the final value for ES. This way the resulting ES

obtained with resampling was 3.29%. This means that under normal market conditions, if the 5% probability event occurs, then the average daily loss will be 3.29%.

The next chart shows VaR and ES for different confidence levels: it can be seen that whereas VaR is both positive and negative, ES is always negative. This is because VaR gives a pointwise estimate of the maximal loss, while ES stems from an averaging of the losses beyond VaR.

Chart 3.: VaR and ES values for the DAX index at different confidence levels (%)



Source: authors' calculations in "R"

In the case of EVaR we applied the resampling method the following way: we calculated the average return and standard deviation for each sample, and then we used the Ahmadi-Javid formula. Lastly the average of these values gave us the final EVaR (3.46% in this case).

3.4 Comparison of the risk measures calculated with the different methods

Let us now compare the risk measures calculated with the three different methods presented in the previous subchapters (see the table below).

	Historical simulation	Monte Carlo simulation	Resampling
VaR (95%)	-2.292%	-2.239%	-2.266%
ES (95%)	-3.293%	-2.796%	-3.290%
EVaR (95%)	3.462%	3.418%	3.460%

Table 3.: Risk measures calculated with the three different methods for the DAX index

Source: authors' calculations in Microsoft Excel
We determined VaR using all three methods. Evidently, the results differ at 95%-os confidence level and daily horizon, due primarily to the fact that the historical data was not normally distributed. The biggest (in absolute value) VaR figure was obtained

for historical simulation (2.29% on a daily horizon and 95% confidence level).

ES and EVaR were also calculated with all three methods. As table 3 shows the values obtained from historical simulation and resampling were very close. The inequality established by Ahmadi-Javid A. (2012) also holds for our data: $EVaR > ES > VaR$. Once again it is evident that EVaR gives the upper bound on the potential losses: daily 3.46% under normal market conditions at a 95% confidence level.

Conclusions

Finally, we must formulate our conclusions, comparing our results with the ones obtained in the literature, and highlighting our contribution to the field of study.

The fact that historical simulation gave us the biggest VaR, ES and EVaR figures is in accordance with the 5-stock portfolio of Ćorkalo S. (2011). Also in line with the other sources, we found bigger values for ES than for VaR (for a given confidence level and holding period). Moreover, similar to Dargiri M. N. et al (2013), the fact that we obtained different values for the three calculation methods, was attributable to the fact that the returns of the DAX index were not normally distributed.

Among the three non-parametric methods (historical simulation, Monte Carlo simulation and resampling), we incline towards the resampling method mainly because we worked with relatively small amounts of historical data, but still were able to generate thousands of samples out of them in order to calculate the risk measures. Another advantage of the resampling method is that it permits the conservation of the properties of the original distribution, the new samples are "grown from the same DNS" as their parent distributions, which in our case means that the small generated samples are also non-normally distributed, rendering them more realistic.

We also showed that the Ahmadi-Javid A. (2012) inequality also holds for the DAX index: $EVaR > ES > VaR$ (for a given confidence level, time horizon and calculation method). The most important feature that arises from this is that EVaR is the upper bound on these risk measures, it is the most conservative and the strictest measure, applying it leads to the most cautious portfolio investment strategies. Therefore EVaR is more suitable for the more risk averse investors (it may be suboptimal for less risk averse ones) or those institutional investors that are constrained by regulation to hold more conservative and less risky positions, such as pension funds, insurance funds etc.

Finally, in this respect we may formulate a recommendation for the Basel Committee on Banking Supervision (BCBS), namely that among the Basel 3 guidelines they should consider the usage of Entropic Value at Risk especially for the institutional investors we've indicated above (thus far there is no mentioning of EVaR among the Basel 3 guidelines).

References

- Ahmadi-Javid, A. (2012). Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications* 155 (3): 1105–1123
- Chernoff, H. (1981). "A Note on an Inequality Involving the Normal Distribution". *Annals of Probability* 9 (3): 533
- Čorkalo, S. (2011), Comparison of Value at Risk approaches on a stock portfolio, *Croatian Operational Research Review (CRORR)*, Vol. 2, p. 81 - 92
- Daniel Zhuoyu Long, Jin Qi (2014): Distributionally robust discrete optimization with Entropic Value-at-Risk, *Operations Research Letters*, Volume 42, Issue 8, December 2014, Pages 532–538
- Dargiri, M.N., Shamsabadi, H.A., Thim, C.K., Rasiah, D. & Sayedy, B. (2013), Value-at-risk and Conditional Value-at-risk Assessment and Accuracy Compliance in Dynamic of Malaysian Industries, *Journal of Applied Sciences*, 13 (7): 974-983
- Iqbal, J., Azher, S. & Ijza, A. (2010). Predictive ability of Value-at-Risk methods: evidence from the Karachi Stock Exchange-100 Index, MPRA Paper, No. 23752
- Manganelli, S. & Engle, R.F. (2001). Value at Risk Models in Finance, Working paper series, European Central Bank, No. 75
- Rudloff, Birgit Jorn Sass, and Ralf Wunderlich (2008): Entropic Risk Constraints for Utility Maximization, working paper, Princeton University
- Zaiwen Wen, Xianhua Peng, Xin Liu, Xiaodi Bai, Xiaoling Sun (2013): Asset Allocation under the Basel Accord Risk Measures, ArXiv working paper
- Zheng Chengli, and Chen Yan (2012): Coherent Risk Measure Based on Relative Entropy, *Applied Mathematics & Information Sciences*.6, No. 2, 233-238

About the authors

Nagy Júlia Tímea: received her Bachelor's degree in Finance and Banking (2012) and her Master's degree in Corporate Financial Management (2014) from "Babes-Bolyai" University, Faculty of Economics and Business Administration, in Cluj Napoca (Romania). Since the 1st of August 2014 she works as a Functional Analyst at Wolters Kluwer Financial Services. Last year, in October 2015, she started her second Bachelor's degree in Business Informatics.

Nagy Bálint Zsolt is an associate professor at Babes-Bolyai University Faculty of Economics and Business Administration since 2013. Since 2010 he is also a senior business engineer at Wolters Kluwer Financial Services. He received his BA and MA from Babes Bolyai University, Romania and his Ph.D. from the University of Pécs, Hungary in 2008.

Juhasz Jacint is a lecturer at Babes-Bolyai University from 2007. He is specialized in Financial Simulations and Corporate Finance having a specific interest in risk management. Juhasz Jacint has also a strong business operation experience being in the management of different Romanian companies in the last 6 years.

TAXATION AND CORPORATE PERFORMANCE: LESS IS MORE

Péter Juhász, Ph.D.¹

Kata Váradi, Ph.D.¹

¹Department of Finance
Corvinus University of Budapest
H-1093, Budapest, Hungary

E-mail: peter.juhasz@uni-corvinus.hu; kata.varadi@uni-corvinus.hu

KEYWORDS

Taxation, firms, growth, optimal tax rate, GDP, simulation.

ABSTRACT

This paper is focusing on how different forms of tax effect the performance of individual companies, the whole economy, and the total tax income of the government. We test fixed, sales-linked and profit taxes under changing circumstances: first we will examine the effect of taxes when the growth rate and the uncertainty is zero, then we will take growth opportunities into account, finally we add uncertainty, too. The main result of this paper are the following. (1) Not only total tax amount but also the form of tax matters. Different types of taxes will influence the business activity in various ways. (2) The extremes are not the best choice: there could be one optimum level of taxation. (3) Increase of nominally fixed taxes near the maximal sustainable level should be lower than the expected growth of the economy; and (4) too high tax burden is more harmful in less stable countries.

LITERATURE REVIEW

The literature on the effect of taxation is quite widespread, since it can be analysed from several aspects. For example the research in this field can be grouped to two main classes, one that are dealing with individuals (eg. Feldstein, 1995, or Surrey, 1970), while the other part is analysing the effect of taxes on corporations (eg. Nielson et al., 2010). This paper will focus only on this later one.

Also this class of research is too broad, since it incorporates questions of the effect of taxation on optimal capital structure (eg. DeAngelo and Masuris, 1980); tax incentives (eg. Graham and Rogers, 2002); tax policy issues (eg. Hall and Jorgeson, 1967). Our research question is more focused, we are not analysing the taxation in general, but we focus on the relation between taxation and corporate performance, and its effect on the whole economy.

Research on this topic states that taxation of corporations has a large impact on the economy as a whole. For example Djankov et al. (2008) states by analysing mid-sized firms, that corporate tax has a

large impact on aggregate investment, foreign direct investment and also on entrepreneurial activity. Also Hall and Jorgeson (1967) found the tax policy had a significant effect on investment behaviour, which cannot be disregarded.

The relation between taxation and company performance, or investment decisions are not as clear, since there are researches that could prove also empirically and theoretically that tax policy is not effective to influence the growth in the long run (Mendoza et al., 1997). While on the other hand it was shown by Levine (1991) that the taxation of financial markets has a large effect on the future growth of the economy.

Greenwood and Huffman (1991) states that government has a crucial role in treating business cycles, since with an effective tax policy the state would be able to stabilize economic cycles. Other papers dealing with the effect of the tax policy on business cycles and growth (eg. Cooley and Hansen, 1992, Mendoza et al., 1994) came to similar conclusions.

Based on the literature taxation has a crucial role in the growth of the economy, and economic cycles. Based on this, the focus of our research is to analyse the effect of corporate taxation of different types on the company performance, and how it effects the economy as a whole in a simple model.

MODEL DESCRIPTION

In our model the behaviour of firms are simulated. All of them are completely equity (E) financed and the only source of additional capital is their own retained profit.

Firms would extend their activity only once the achieved after-tax profit over equity (ROE – return on equity) is higher or equal to the required rate of return of the owners (rE). If return is less than required, owners are not satisfied and usually are hesitating to reinvest the profit. Of course they know that fluctuations in market and efficiency measures are normal.

This is why, in our model shareholders consider the three year average ROE and compare that to the required rate. Once being above that level they intend to increase the business as that generates positive net present value (NPV).

If the given firms are willing to grow they will set the amount of reinvestment in line with the expected growth of the market. The market is dominated by the firms included in the simulation so in period t they compare their expected growth rate (g') in period t-1 (that is a prediction for this given period change) for the local market with the percentage increase of the total local market sales (g) of the companies in the model. It is the average of those two quantities that firms will use as a prediction for the next period according to Equation 1:

$$g_{Local}'_t = \frac{g_{Local}'_{t-1} + g_{Local}_t}{2} \quad (1)$$

At the start of the simulation, $g'0$ (expected growth for next period) is given as a parameter.

If the company is profitable, but average ROE falls below the required, owners withdraw all earnings as dividend. Once the firm is in the red, loss will decrease the amount of equity that is equal to the total invested capital (IC) as no debt is used in the model. (In real life unhappy owners can even decrease invested capital further by selling assets and repurchasing shares, but in this model they cannot do so.) Once its equity reaches zero, the firm stops its operation.

In any period the operation of the firm is simulated as follows. Based on the start of period E using a pre-set Sales/IC (asset turnover) ratio Sales is calculated. Costs of manufacturing is described by the Labour cost/Sales and Material cost/Sales ratios. All three of these ratios follow a normal distribution with given average and standard deviation.

Beside these expenses, different taxes are also deducted from sales. The so calculated operational profit (earnings before interest and taxes – EBIT) is equal to profit before tax (PBT) as no debt is used. If PBT is positive, the firm has to pay a proportional corporate tax, and the leftover quantity is called profit after tax (PAT). Dividing that by the start of period equity we get return on equity we use for to decide whether owners want to grow the business further.

In this model we included three different taxes:

(1) Firms may be required to pay fixed amount taxes that may grow independently from the business activity of the companies. Some real life example on that could be yearly fees of public registration, statistical reports, publishing financial statements, or costs of legal actions required by the state. In our model, this nominally fixed tax is increased by steady percentage (bigger or equal to zero) each period.

(2) Some taxes are proportional to the business activity but not its profitability. Those are modelled by including a tax on sales. Real life examples include special fees imposed on trucks for using public highways (proportional to the distance used), environmental duties on packaging materials (proportional to the quantity used), or state authority supervisory fees linked to production or sale of certain

goods (e.g. food, gasoline, livestock) There is also an explicit sales tax of 2 percent applied in Hungary.

(3) Taxes on profit are paid by firms generating a PAT more than zero. In our model a fix percentage charge is applied, while in many countries you may see different tax rates applied for SMEs and bigger firms or offering corporate tax reduction to some industries or to firms performing specific activities like huge investments or employing handicapped. Firms with a PAT lower than zero pay no tax but contrary to what is common in many countries in the model these losses cannot be used to decrease tax base in the following years.

In this model, all firms are under the same tax regime so tax rates do not differ across them.

Our model simulates the behaviour of individual firms that might be different both in productivity (Sales/IC) and efficiency (Labour expenditure/Sales, Material expenditure/Sales). For each of these periods the total amount of sales, and tax collected are recorded together with total added value generated (Sales-Material expenses), called as GDP at macro level. We also keep a records of how expected and realised market growth rates developed.

COMPARING TAX TYPES IN A NON-GROWTH ECONOMY

As presented in the model description part, we have three types of taxes in this simulation. We may sum up the effect of those on ROE by the following formulas of Equation 2 and Equation 3.

$$ROE = \frac{PAT}{Equity} = \frac{IC}{Equity} * \frac{Sales}{IC} * \quad (2)$$

$$* \frac{EBIT}{Sales} * \frac{PBT}{EBIT} * \frac{PBT * (1 - corp. tax)}{PBT}$$

$$EBIT = Sales * \quad (3)$$

$$* \left(1 - \frac{Sales tax}{Sales} - \frac{Material exp.}{Sales} - \frac{Labour}{Sales} - Fixed tax \right)$$

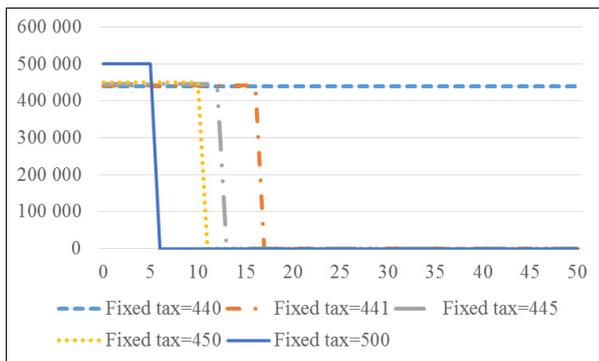
To compare their effect on the economy we set up a hypothetical country with firms only focusing on a single market, and use no foreign capital. Start-up parameters are shown in Table 1.

First, to eliminate fluctuations, standard deviations are set to zero just as expected growth for the markets. In that case, firms do not have the opportunity to grow, so fixed tax creates a constant inflow for the state and can be increased up to the level of PAT without all taxes (in this case 440 units).

Table 1. Start-up parameters for comparing effects of taxes

N of firms	1000
Equity/firm	1000
IC/Equity	100,00%
r_E	10,00%
Sales/IC average	110,00%
Sales/IC std. dev.	0,00%
Labour exp./Sales average	20,00%
Labour exp./Sales std. dev.	0,00%
Material exp./Sales average	40,00%
Material exp./Sales std. dev.	0,00%
Local Sales/Sales	100,00%

Above that the equity decreases due to the negative PAT and companies soon stop operating as it can be seen in Figure 1.



Figures 1: Tax income across periods at different fixed tax levels

All other forms of taxes reduce ROE and once pushing that below zero the destroying of the tax base starts. For the given case any sales tax above 40 percent will have similar effect, but the destruction will show somewhat different pattern according to Figure 2.

We get very similar figures when charting for different levels of corporate tax. For corporate tax at all possible inputs the critical level is 100 percent.

Of course political decision makers should not focus on maximising state income rather than (at least within the frame of this model) on maximising GDP (total added value). Note that for the previous cases as growth was not possible GDP remained the same until reaching the destruction level of the given tax, and taxation was only about redistributing GDP. (Tax decreases income of shareholders and boost the income of those receiving governmental transfers.)

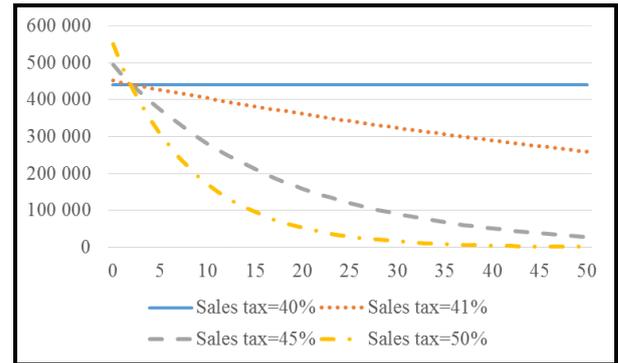


Figure 2: Tax income across periods at different sales tax levels

TAX AND GROWTH

Now, let us add a 5 percent expected growth opportunity on the local market. In this case firms would reinvest a part of their profit once $ROE \geq r_E$. For the shake of this model r_E equals 10 percent.

The no-tax ROE is 44 percent, which is more than enough to finance a 5 percent growth. So any tax system that keeps after-tax ROE above 10 percent, would keep maximum growth and not jeopardize the long term existence of the firm. (The critical value for that is 340 in case of fixed tax.) After-tax ROEs between 0 and 10 percent end up with a zero growth economy, while negative shareholder return will destroy the companies.

Growth adds also another dimension to the optimal taxation problem. Fixed tax may mean an increasing or decreasing burden for firms depending on whether the growth of the firm is higher or lower than that of the tax amount. Figure 3 illustrates the effect of fixed tax starting from 340 and growing at different rates.

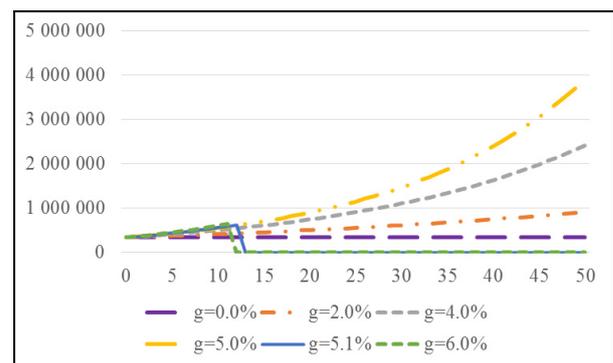


Figure 3: Tax income across periods at different fixed tax growth levels

As here taxation only limits economic growth when fixed amount grows faster than the firms themselves, GDP is maximised with any tax growth rate not higher than 5 percent. The exact value below that level is just a decision about redistribution.

Still, we have to notice that there could be a potential pitfall of taxation once the start-up level of fixed tax is less than the maximum (e.g. here 340). For some years even with growth rates above 5 percent the economy grows, but ROE decreases. So at a certain point, the whole systems collapses all of a sudden according to Figure 4.

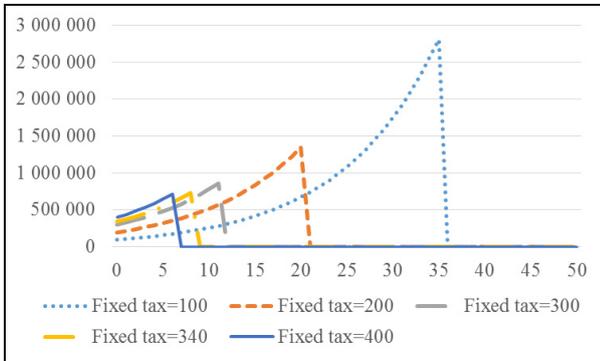


Figure 4: Tax income across periods at fixed tax levels with a yearly growth of 10 percent

The lower the initial level the later this happens so in a real world a government may go on for years with unsustainable taxation before noticing. When focusing on GDP we may have a chance to track the problem earlier, as Figure 5 shows.

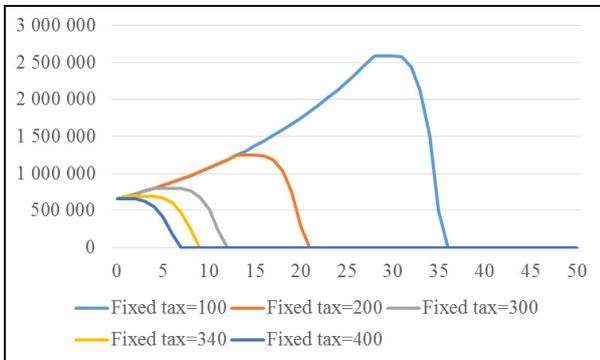


Figure 5: GDP across periods at fixed tax levels with a yearly growth of 10 percent

As other kind of taxes use quantities indexed to growth as tax base, we have no such a kind of problem: if rates are not too high at start-up, the economy will not collapse. Putting it differently: these kind of taxes do not allow politicians to overtax firms and stay unnoticed for years. That might be a reason for decision makers suffering from myopia to prefer duties not proportional to the business activity rather manually indexed.

Until now in all our examples r_E was higher than the maximal growth rate available on the market (g) due to which we either have seen companies growing at the

maximum possible rate ($ROE \geq r_E$) or not at all, or sometimes even decreasing in sales if losing their equity. The problem is slightly more complex once the market growth rate is higher than r_E . Under this condition depending on the ROE achieved firms may grow at any rate between g and r_E . So fine tuning of the tax system (not applying maximum charge) has a radical effect on the performance of the economy.

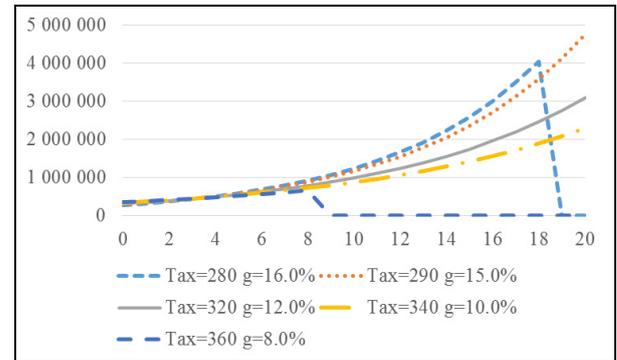


Figure 6: Tax income across periods at different fixed tax levels and yearly growth rates

Imagine the market offers a growth opportunity of 15 percent per period. Figure 6 demonstrates how more moderate fix tax amounts end up with far higher total tax income over time. (To assure maximum growth fixed tax may not be higher than 290 at initiation and to be sustainable a higher amount can only grow at lower rate.) Any duty above 340 would immediately decrease ROE below r_E , while any growth rate above 15 percent will sooner or later make the given tax amount to push future ROE below future r_E

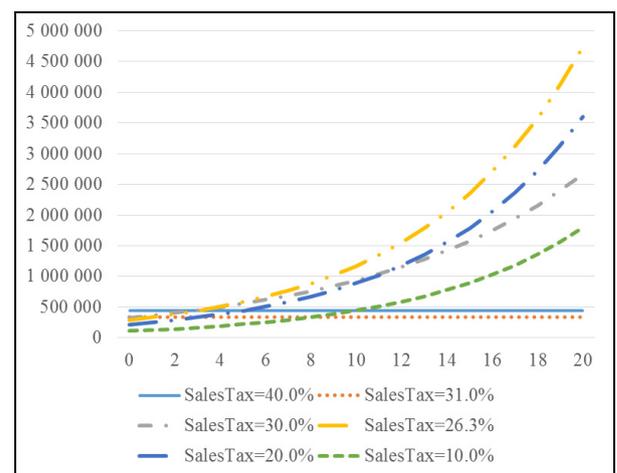


Figure 7: Tax income across periods at different sales tax levels

When considering sales tax, any rate above approximately 26.3 percent would leave less than 15 percent ROE with the firm, so it cannot use the total growth potential offered by the market, while any rates above 40 percent would make the companies to lose money. It is at around 31 percent that ROE equals r_E , and firms stop to grow.

As seen on Figure 7 we have here the same problem as in case of fixed duties: it is not the long term optimum tax rate that would provide the highest income in the short run.

As for corporate tax we see similar patterns (Figure 8). The two critical values are 65.8 percent (leaving enough ROE to use the total growth opportunity) and 77.3 percent (lowering ROE to 10 percent, blocking growth completely). Of course above 100 percent, due to the continuous decrease of equity the economy will collapse.

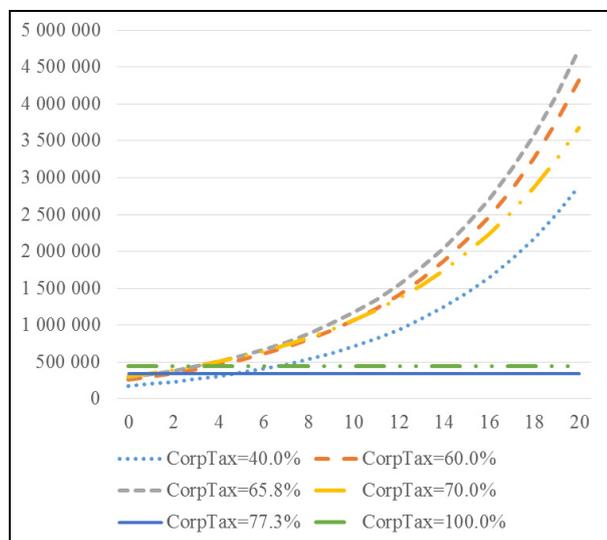


Figure 8: Tax income across periods at different corporate tax levels

UNCERTAINTY AND TAXES

As a next step we introduce some uncertainty in to the model. Three ratios: Sales/IC, Material expenses/Sales, and Labour expenses/Sales follow a normal distribution with the same expected values used until now. Relative standard deviation (std. dev./average) was set to 2 percent. Simulation is done at firm level, so the expected growth of the whole economy is 15 percent and more smooth than that of the individual firms.

Without uncertainty the critical values were 290 ($ROE=g$) and 340 ($ROE=r_E$). But once we add uncertainty, applying the former with 15 percent yearly growth will not lead to a sustainable economy anymore. If individual performance of a firm falls short of the expected, it will be unable to grow because of the tax reducing ROE below the required level. What is more, one single event like that will kill the given firm sooner or later as the tax keeps on increasing at the

same rate while the company will not grow in that given year, so the tax charge for the firm will be higher next year. But market conditions will not allow to boost growth above the expected level so it will be impossible to push back the tax charge to the original level. After some unsuccessful periods, ROE will have nearly no chance to reach r_E . Due to this phenomenon only a taxation system with a tax amount growth far lower than that of the economy would be sustainable as illustrated on Figure 9. Note that for a 50 period time interval only a tax of 265 could grow at 15 percent – an amount 8.6 percent less than the theoretical maximum but still providing the maximum tax income over the long run.

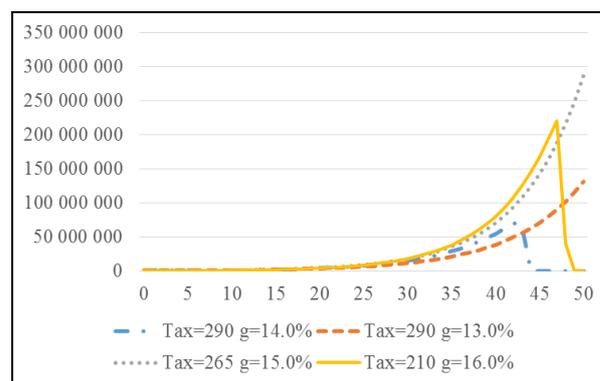


Figure 9: Tax income across periods at fixed tax levels with different yearly growth rates with uncertainty

Figure 9 demonstrates that tax systems generating higher income in the short run may be less effective in the long run. The best long term option in our example produces less state income in each of the first five years than a tax of 290 with any growth rate between 13 and 30 (!) percent. So the time horizon using which politicians optimise their decisions (e.g. the length of government term) may have a dramatic effect on the probability of introducing optimal taxation schemes.

Considering Sales tax, the critical value without uncertainty was 31 percent reducing ROE to the minimum required level and 26.3 to keep maximum growth potential. In Figure 10 under uncertainty rates not allowing room for growth slowly destroy the economy (lowering GDP) while tax income stagnates. Not using the maximum tax rate to leave room for performance fluctuations and keep maximum growth pays off in the long run. A rate of 24 percent even generates more income than the theoretical maximum rate of 26.3 percent from the 13th year on.

Finally, let us take a look on how different corporate tax rate perform under uncertainty. Previously, critical values were 65.8 and 77.63 percent. Under the given level of uncertainty even a rate of 60 percent performs better than that of 65.8 percent, and 63 percent over-performs both of those in Figure 11.

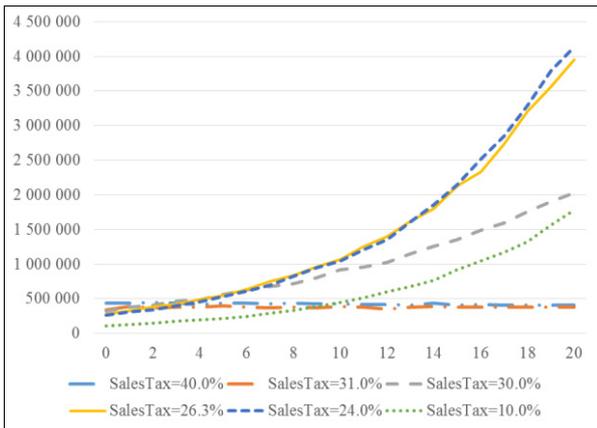


Figure 10: Tax income across periods at various sales tax levels under uncertainty

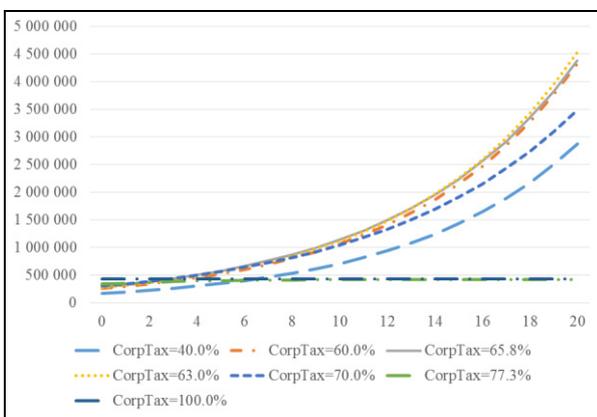


Figure 11: Tax income across periods at different corporate tax levels under uncertainty

It is worth noting that in the long run (after period 10) the tax income produced by the rate of 77.3 percent is only by 3 percent less than that with a 100 percent rate while GDP is 20 percent higher. That is due to the 50 percent probability of ending up with a rate above the required and so with at least a slight growth. Actually a rate of 77.25 percent already produces more state income from period 12 on while the GDP is higher in any period after the second adding already more than 30 percent from period 12 on.

CONCLUSION

For this paper we modelled the growth decisions of individual firms that could be hit by two types of tax: one of them being nominally fixed and increased at a steady rate, the other being proportional to sales (under model conditions similar to classic corporate tax on profit). Using our simple model we may draw at least four very important conclusions.

(1) Our simulation has highlighted that different types of taxes will influence the business activity in various ways. When politicians decrease one rate and increase another to keep budget balanced it is far more than

redistribution and may make a huge difference even if the current total tax collected remains the same.

(2) We also showed that there could be one optimum level of taxation, though without using the classic argumentation about high rates increasing tax evasion or decreasing willingness to start new business. While higher than optimum tax rates may generate extra income for the state in the short run, due to cutting back on growth rates the whole economy will suffer in the long run.

(3) When deciding about the increase of nominally fixed taxes near the maximal sustainable level (this may happen even at a given combination of different taxes), decision makers have to use a considerably lower index rate than the expected growth of the economy, a result that might be counterintuitive. Due to uncertainty regarding e.g. efficiency, material costs, and labour prices highly taxed firms may not have enough earning to retain to take profit of the market growth. In such a case sooner or later the ever growing tax will ruin them.

(4) Generally, in an economy with companies of more fluctuating performance government should leave more room for efficiency fall-backs when setting tax rates of any kind to maximise economic growth (and tax income) in the long run. In other words too high tax burden is more harmful in less stable countries.

Further research opportunities include introducing other types of taxes in the model, simulating an economy with firms of different productivity and efficiency, various leverage rates, or facing differences in growth opportunities due to focusing on different markets.

LIMITATIONS

In the real economies there is much more room for a tailor made taxation system than offered in this model. Taxes applied were very simple and only used one kind of tax with flat rate for all firms. In contrast to that, most tax schemes use several rates, tax base reduction options, and allowances, e.g. some duties and taxes may not be to be paid by firms under a given size, in specific industries or operating in underdeveloped areas.

The effect of taxation on the number of newly established companies was not considered and in our model there was no other exit for the owners but to wait until their equity completely evaporated. In real life owners would liquidate investments that are not likely to be profitable enough ($ROE \geq r_E$) in the future (GDP would decrease sooner), and in such an industry none would start new business either. Of course also decision maker may notice their mistakes earlier than the complete collapse of the economy, also thanks to macro analysts, academic researchers or the protest of entrepreneurs and business people.

REFERENCES

- Cooley, T.F. and Hansen, G.D. (1992): Tax distortions in a neoclassical monetary economy. *Journal of Economic Theory*, 58(2), pp.290-316.
- DeAngelo, H. and Masulis, R.W. (1980): Optimal capital structure under corporate and personal taxation. *Journal of financial economics*, 8(1), pp.3-29.
- Djankov, S., Ganser, T., McLiesh, C., Ramalho, R. and Shleifer, A. (2008): *The effect of corporate taxes on investment and entrepreneurship* (No. w13756). National Bureau of Economic Research.
- Feldstein, M. (1995): The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act. *Journal of Political Economy*, pp.551-572.
- Graham, J.R. and Rogers, D.A. (2002): Do firms hedge in response to tax incentives?. *The Journal of finance*, 57(2), pp.815-839.
- Greenwood, J. and Huffman, G.W. (1991): Tax analysis in a real-business-cycle model: On measuring Harberger triangles and Okun gaps. *Journal of Monetary Economics*, 27(2), pp.167-190.
- Hall, R.E. and Jorgenson, D.W. (1967): Tax policy and investment behavior. *The American Economic Review*, 57(3), pp.391-414.
- Levine, R. (1991): Stock markets, growth, and tax policy. *The Journal of Finance*, 46(4), pp.1445-1465.
- Mendoza, E.G., Milesi-Ferretti, G.M. and Asea, P., 1997. On the ineffectiveness of tax policy in altering long-run growth: Harberger's superneutrality conjecture. *Journal of Public Economics*, 66(1), pp.99-126.
- Mendoza, E.G., Razin, A. and Tesar, L.L. (1994): Effective tax rates in macroeconomics: Cross-country estimates of tax rates on factor incomes and consumption. *Journal of Monetary Economics*, 34(3), pp.297-323.
- Nielsen, S.B., Raimondos-Møller, P. and Schjelderup, G. (2010): Company taxation and tax spillovers: separate accounting versus formula apportionment. *European Economic Review*, 54(1), pp.121-132.
- Surrey, S.S. (1970): Tax incentives as a device for implementing government policy: A comparison with direct government expenditures. *Harvard Law Review*, pp.705-738.

AUTHOR BIOGRAPHIES

PÉTER JUHÁSZ is an Associate Professor of the Department of Finance at Corvinus University of Budapest (CUB). He holds a PhD from CUB and his research topics include business valuation, financial modelling, and performance analysis. His e-mail address is: peter.juhasz@uni-corvinus.hu

KATA VÁRADI is 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, bonds markets and capital structure of companies. Her e-mail address is: kata.varadi@uni-corvinus.hu

MENTAL FRAMING IN RISK-AVERSION DYNAMICS

AN EMPIRICAL INVESTIGATION OF INTERTEMPORAL CHOICE

Mihály Ormos
Dusán Timotity
Department of Finance,
Budapest University of Technology and Economics
Magyar tudosok krt. 2., 1117 Budapest, Hungary
ormos@finance.bme.hu and timotity@finance.bme.hu

KEYWORDS

Asymmetric volatility; Risk seeking; Prospect theory; TGARCH; Volatility dynamics

ABSTRACT

This paper provides an empirical investigation of the mental framing based explanation for heteroscedasticity by Ormos and Timotity. We find empirical support for their model from two different point of view: first, the analysis of a huge individual trading dataset shows that investors indeed become risk-seeking right after losses and more risk-averse subsequent to gains; second, the parameter estimation of our volatility model yields the predicted negative relationship between abnormal returns and subsequent volatility.

INTRODUCTION

Time-varying volatility (heteroscedasticity) of asset returns has attracted much research in the recent decades. Since the milestone papers of Engle (1982) and Bollerslev (1986) a great number of scholarly paper has been devoted to the topic. Their findings indicate that the phenomenon can be modeled by GARCH type models; however, an important aspect of the autoregression puzzle, the asymmetry in the volatility process still misses a robust explanation with empirical investigation. We aim to fill this void in the literature by providing empirical results for the theoretical model of Ormos and Timotity (2015), henceforth OT.

First, we show that, in line with results of Thaler and Johnson (1990), investors become risk-seeking following losses and risk-averse subsequent to gains if the opportunity of breaking-even is included in the choice set, which, in fact, almost always applies to asset returns. According to the model of OT, this pattern is due to the intertemporal mental framing of investors, which causes a negative relationship between previous unanticipated outcomes and risk-seeking. We confirm their hypothesis by analysing a large dataset containing individual trades and portfolio allocations.

Second, we present that the individually measured patterns of risk-aversion apply at the market level as well. Here, we find empirical support for the proposed theoretical volatility model of OT and confirm the existence of a negative relationship between previous market shocks and subsequent asset price volatility.

The paper is structured as follows: in section 2.1 the patterns in investors' intertemporal choice are discussed, then in 2.2 the volatility model is estimated. Finally, in section 3 we provide a brief conclusion on the main results.

EMPIRICAL RESULTS

In this section we present our empirical results supporting the theoretical model of OT in two different ways: first, investors' dynamic behavior is tested on a large sample containing individual trading data; second, an empirical parameter estimation of our volatility model is provided using CRSP database consisting of the daily log-returns of the Standard and Poor's 500 index member listed on 10 September, 2014. The analysed period covers 21 years from 10 September 1993 to 10 September 2014.

Patterns in intertemporal choice

We empirically investigate whether losses and gains induce risk-seeking and more risk-averse behavior respectively. As OT's theoretical model argue, this behavior is a response to loss-aversion in a dynamic context, that is, investors are reluctant to realize losses (either physically or mentally) and try to break even in order to obtain their initial benchmark on average. According to equilibrium asset pricing, higher required return that compensates for the previous loss can only be reached by investing in assets with increased risk; therefore, combined with the change in risk attitude, losses increase the volatility of returns in the subsequent period. Gains follow the opposite pattern: investors fear of losing the previous wealth, hence, they invest into less risky portfolios since the initial benchmark level is still reachable with the latter.

The data and methodology of this analysis are as follows: Our sample is similar to that of Barber and Odean (2000) consisting of the transactions and descriptive data of 158,006 accounts at a large discount brokerage firm from January 1991 to December 1996. In this paper we aim at defining the change in the riskiness (as measured by volatility) of investors' portfolio; therefore, only common stocks investments are considered, since a meaningful amount of historical returns and realized volatility can only be calculated for these assets. Nevertheless, findings in this reduced sub-

sample should be representative for the whole sample as the former account for 64% of the latter as measured by the number of observations. Altogether, the dataset containing at least one common stock transaction in the period includes 104,225 accounts, which can be further decomposed based on the type of the account, in which we apply cash, IRA and margin accounts as control variables, and the equity held by the related household at the end of the period. In Table 1 the descriptive statistics of these sub-samples are presented.

Table 1: Descriptive statistics of the sample

	All accounts	Cash accounts	IRA accounts	Margin accounts
Num. of accounts	104,225	22,995	37,155	10,328
Mean equity	68,293	39,859	48,988	47,953
Median equity	18,288	8,419	21,549	4,426
St. dev. of equity	300,450	129,257	129,017	247,607
Num. of trades	1,969,747	260,039	486,889	255,759
Mean number of trades	19	11	13	25

Notes: The table shows the descriptive statistics of the trading accounts included in our dataset.

In return calculations we use different types of mental frames. First, we assume that when selling occurs the profit is measured as the selling price relative to the pre-transaction average buy price of an asset. However, as the long position in an asset may include numerous buy transactions before selling the stock, we argue that if the representativity or anchoring heuristics are responsible for the change in the risk attitude, the most recent information (i.e. the price of the last buy transaction) is the main factor in utility perception. Having calculated the gain or loss, the asset into which the realized money flows in the subsequent buy transaction is defined. Related to both the bought and sold assets the variance and standard deviation of daily returns in the preceding year are calculated. Finally, based on the aforementioned parameters, regressions are estimated to analyse whether the risk of the targeted asset is driven by the previous outcome.

As the number of trades of separate investors is often too small to capture individual account effects, we apply a pooled data structure. Furthermore, since the number of accounts and trade numbers justify the use of the central limit theorem, our regressions are based on OLS estimations.

The first regression (first 2 columns in Table 2) applies a simple estimation of the variance of the targeted asset including the profit (the return based on the average buy price) of the previous transaction as the independent variable, that is

$$\sigma_{b,i}^2 = \hat{\alpha} + \hat{\beta}_1 \bar{r}_{s,i} + e_i, \quad (1)$$

where $\sigma_{b,i}^2$ and $\bar{r}_{s,i}$ stand for the variance of the asset in the subsequent buy transaction and the average return of

the realized sell transaction of each i trade pair respectively.

In the second regression we test whether the change in the definition of the return increases significance and goodness-of-fit. This estimation is shown in Eq. (2) where the previous profit $r_{s,i}$ is measured as the return on the price of the last transaction.

$$\sigma_{b,i}^2 = \hat{\alpha} + \hat{\beta}_1 r_{s,i} + e_i, \quad (2)$$

One may argue that the variance also correlates with the risk of the sold asset as well: an investor may have a preference for risky assets, which could lead to a biased estimation of $\hat{\beta}_1$ in the previous equation. Therefore, the third regression (Eq. (3)) includes $\sigma_{s,i}^2$ as the variance of the sold asset using the return on the last buy price respectively.

$$\sigma_{b,i}^2 = \hat{\alpha} + \hat{\beta}_1 r_{s,i} + \hat{\beta}_2 \sigma_{s,i}^2 + e_i, \quad (3)$$

According to equilibrium pricing, investors do require a premium for risk; thus, their expected return is different from zero. Including this finding in the fourth regression, a new definition of return may provide a better fit to utility perception: here the perceived return is defined as the deviation from the historical (one year) expected return at the last buy transaction preceding the sell transaction of an asset. In other words, we assume that investors form their non-zero expectations at the time they invest into an asset based on its performance in the past. Accordingly, as both the length of time between last buy and subsequent sell transactions and the risk of assets varies throughout the data, another adjustment is required: the expected return is not the same for each transaction, hence, we standardize the deviation from the expected return by dividing it by the number of days between the buy and sell transactions. Subsequent to this definition we use this daily average deviation from the expectation as an independent variable as in the following Eq. (4), where t_s and t_{pb} stand for the time when the sell and the previous buy transactions occurred:

$$\sigma_{b,i}^2 = \hat{\alpha} + \hat{\beta}_1 r_{std,s,i} + \hat{\beta}_2 \sigma_{s,i}^2 + e_i : r_{std,s,i} = \frac{r_{s,i} - E(r_i | t = t_{pb})}{t_s - t_{pb}} \quad (4)$$

In order to be able to distinguish effects of previous gains from losses we apply two separated variables in regression five as defined in Eq. (5):

$$\sigma_{b,i}^2 = \hat{\alpha} + \hat{\beta}_1 r_{-std,s,i} + \hat{\beta}_2 r_{+std,s,i} + \hat{\beta}_3 \sigma_{s,i}^2 + e_i : r_{-std,s,i} = \min(r_{std,s,i}, 0), r_{+std,s,i} = \max(r_{std,s,i}, 0) \quad (5)$$

Having analysed the effects of previous outcomes on risk attitude as measured by variance, we provide further tests that include volatility instead of the former. The importance of this additional analysis is already highlighted, where we discussed that asset prices in prospect theory are driven by standard deviation rather than variance. Hence, in further regressions we apply volatility as the dependent variable. The sixth regression

is the same as Eq. (5) except for the previously defined change in the definition of risk.

Our extensive dataset covers further parameters related to each trading account; in particular, the equity held at the end of the period and the type of the account is included as well. In further regressions we also apply these latter measures as control variables and investigate differences between the subgroups. The seventh regression is defined as in Eq. (6), where $E_i, D_{C,i}, D_{I,i}$ and $D_{M,i}$ stand for the equity, the cash type dummy, the IRA type dummy and the margin dummy of the account related to the i^{th} transaction respectively.

$$\sigma_{b,i} = \hat{\alpha} + \hat{\beta}_1 r_{std,s,i} + \hat{\beta}_2 \sigma_{s,i} + \hat{\beta}_3 E_i + \hat{\beta}_4 D_{C,i} + \hat{\beta}_5 D_{I,i} + \hat{\beta}_6 D_{M,i} + e_i \quad (6)$$

In regression eight we modify Eq. (6) according to Eq. (5), that is, by separately estimating the coefficients of gains and losses. Then, in subsequent estimations we apply this latter frame in subgroup estimations: in the ninth equation the effects for accounts with equity value above its median (i.e. the top 50% of investors ranked by equity value) are estimated, whereas the tenth calculates coefficients for the bottom 50%. In the last three regressions effects for subgroups with a cash, IRA and margin account types are estimated.

Table 2: Regression results

Panel A						
	Subsequent σ^2 (Eq. 1)		Subsequent σ^2 (Eq. 2)		Subsequent σ^2 (Eq. 3)	
	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	2.32E-03	0.0000	2.32E-03	0.0000	2.22E-03	0.0000
Average return	-8.60E-05	0.0010	-	-	-	-
Return on the last trade	-	-	-9.47E-05	0.0005	-1.09E-05	0.6885
Previous variance	-	-	-	-	4.94E-02	0.0000
Adjusted R-squared	0.0000	-	0.0000	-	0.0026	-

Panel B						
	Subsequent σ^2 (Eq. 4)		Subsequent σ^2 (Eq. 5)		Subsequent σ	
	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	2.21E-03	0.0000	2.17E-03	0.0000	3.00E-02	0.0000
Previous variance	2.21E-03	0.0000	4.69E-02	0.0000	-	-
Expected return	-	-	-	-	-	-
Difference of last return	-1.63E-03	0.0003	-	-	-	-
Positive diff. of last return	-	-	4.72E-03	0.0000	6.65E-03	0.0071
Negative diff. of last return	-	-	-7.78E-03	0.0000	-1.48E-02	0.0000
Previous volatility	-	-	-	-	2.32E-01	0.0000
Adjusted R-squared	0.0027	-	0.0031	-	0.0386	-

Panel C								
	Subsequent σ (Eq. 6)		Subsequent σ		Subsequent σ if Equity \geq Median		Subsequent σ if Equity $<$ Median	
	Coef	p-value	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	3.08E-02	0	3.08E-02	0.0000	3.09E-02	0.0000	3.16E-02	0.0000
Difference of last return	-4.11E-03	0.0135	-	-	-	-	-	-
Positive diff. of last return	-	-	5.53E-03	0.0251	7.49E-03	0.0429	2.43E-03	0.4627
Negative diff. of last return	-	-	-1.35E-02	0.0000	-9.93E-03	0.0098	-1.47E-02	0.0000
Previous volatility	2.30E-01	0	2.28E-01	0.0000	1.99E-01	0.0000	2.48E-01	0.0000
Equity	-2.06E-09	0	-2.06E-09	0.0000	-1.61E-09	0.0000	-5.31E-08	0.0000
Cash dummy	-5.33E-04	0.0008	-5.25E-04	0.0010	-5.60E-04	0.0213	-9.45E-04	0.0000
IRA dummy	-1.66E-03	0	-1.67E-03	0.0000	-3.16E-03	0.0000	-1.96E-04	0.2380
Margin dummy	1.50E-03	0	1.49E-03	0.0000	2.57E-03	0.0000	1.37E-04	0.4620
Adjusted R-squared	0.0419	-	0.0419	-	0.0348	-	0.0467	-

Panel D						
	Subsequent σ for cash account		Subsequent σ for IRA account		Subsequent σ for margin account	
	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	3.00E-02	0.0000	2.84E-02	0.0000	3.06E-02	0.0000
Positive diff. of last return	5.11E-02	0.0000	7.26E-03	0.1573	-1.28E-02	0.0388
Negative diff. of last return	-9.60E-04	0.9195	-3.50E-03	0.4974	-1.77E-02	0.0009
Previous volatility	2.44E-01	0.0000	2.60E-01	0.0000	2.66E-01	0.0000
Equity	-6.65E-09	0.0000	-4.80E-09	0.0000	-9.63E-10	0.0000
Adjusted R-squared	0.0531	-	0.0565	-	0.0479	-

Notes: The table represents regression results for equations one to six and their modified versions in Panels A, B, C, D. The dependent variables are listed in the columns. The Coef columns represent the estimated coefficients for the parameters listed in the rows, whereas the p-value columns stand for the probability of an incorrect rejection of the zero null hypothesis.

In Table 2 we provide the empirical results of the estimations: results for groups of regressions one to six and their modified versions in Panel A, B, C and D. Results of the first four regressions indicate that regardless of the type of return, the aggregate effect of previous outcomes on risk attitude is significantly negative even if the previous variance is included, which supports the theory of dynamic loss-aversion. Even though, we find a minor increase in the significance by changing the reference point from the average return to the return relative to the price of the last buy transaction first, then to the return relative to the historical expected return second, the extremely low adjusted R-squared values indicate non-linear dynamics or missing variables. Regression five yields a possible

reason for this latter finding: gains and losses have a distinct effect on risk attitude, although, separating the previous outcomes by their sign does add a lot to the goodness-of-fit of the latter models.

This problem is well handled by changing the risk measure to volatility: the sixth regression shows that the adjusted R-squared value jumps.

Results of the volatility estimation of seventh regression indicate four main findings: first, the aggregate effect of previous outcomes is significantly negative again; second, equity has negative effect on risk-appetite indicating that investors holding larger amounts in capital assets invest into less risky portfolios; third, market participants with cash and retirement (IRA) accounts also avoid risk shown by their negative coefficient; fourth, margin account holders have higher appetite for risk as shown by the positive relationship between subsequent volatility and the margin dummy.

Altogether, regressions in Panel C all indicate a similar pattern as before: negative differences relative to the expected return have a significant and negative effect on the subsequent risk-appetite, whereas positive differences are either much less significant or not significant at all. In particular, regressions nine and ten show that choices of high-income investors are just as sensitive to previous outcomes as low-income investors. In Panel D regression results show a somewhat mixed picture: although coefficients are not significant everywhere, the previous patterns apply to every subgroup except for the coefficient of the positive previous return of margin account holders. In this latter group, both previous gains and losses are significantly negative leading to lower and higher subsequent volatility respectively.

Altogether, we find similar results to the aggregated regression of Eq. (6) and its adjustment for separated gains and losses. Although, for positive deviations from the expected return we find a statistically significant positive effect on subsequent volatility, we argue that the low p-values are due to the extremely high number of observations. According to OT, positive deviations from the expected return are also negatively correlated with subsequent volatility; nevertheless, since volatility is non-negative, huge realized gains lead to exactly the same portfolio choice (i.e. the risk-free asset) as a gain that is just high enough to cover two subsequent periods of the required return. Therefore, positive returns higher than a relatively small level (at least twice of the expected return) cannot be described by a linear relationship with volatility but are driven by a random process. This leads to the fact that for a reasonable number of observations, where the case of “too big to fail” does not apply, p-values of the positive coefficient should not indicate a significant effect. The last three regressions in Panel C (eighth to tenth regressions), in which the p-value of the coefficient of previous gains is much higher than that of losses, suggest such relationship; however, for such high number of observations a tiny effect may prove to be significant.

We argue that this effect may be due to a non linear relationship between previous gains and volatility.

A methodologically solid way to handle this non-linearity would be to use a simple dummy variable for positive shocks. The intuition behind this idea is that if the expected return is relatively very small compared to the positive shocks, then, shocks exceeding this expected return have a constant effect on volatility, since investors would not and cannot reduce their required return and portfolio volatility to values below zero: they hold assets providing at least the risk-free return with zero volatility. Therefore, there is a discontinuity in the model for gains, which can be handled with the use of a dummy variable.

In the followings, we compare the results of the aforementioned model applying a dummy variable for gains and the model assuming a linear relationship between previous gains and subsequent volatility. Table 3 represents our findings.

Table 3: Regression results of volatility dynamics

	Subsequent σ			
	Coef	p-value	Coef	p-value
(Intercept)	3.08E-02	0.0000	3.11E-02	0.0000
Positive diff. dummy	-	-	-6.12E-04	0.0000
Positive diff. of last return	5.53E-03	0.0251	-	-
Negative diff. of last return	-1.35E-02	0.0000	-9.73E-03	0.0001
Previous volatility	2.28E-01	0.0000	2.29E-01	0.0000
Equity	-2.06E-09	0.0000	-2.07E-09	0.0000
Cash dummy	-5.25E-04	0.0010	-5.49E-04	0.0006
IRA dummy	-1.67E-03	0.0000	-1.67E-03	0.0000
Margin dummy	1.49E-03	0.0000	1.48E-03	0.0000
Adjusted R-squared	0.0419	-	0.0420	-

Notes: The table represents regression results for two regressions between previous outcomes and subsequent volatility. The dependent variable is listed in the columns, the Coef columns represent the estimated coefficients for the independent variables listed in the rows, whereas the p-value columns stand for the probability of an incorrect rejection of the zero null hypothesis.

The results indicate three important findings: first, by avoiding the discontinuity problem the regression model support our idea of a positive relationship between previous gains and volatility instead of linearity; second, this relationship becomes much more significant than in the linear model and therefore, all the variables have extremely low p-values; third, the adjusted R-squared also increases in the new model suggesting a better fit with the dummy variable. Hence, altogether the findings support the negative relationship proposed in the theoretical model.

In conclusion, we argue that the results presented in this subsection confirm the empirical validity of the behavioral side of our explanation. The aggregate coefficient of previous outcomes is negative and significant everywhere, even in regressions where other control variables are included. In particular, it seems irrelevant whether we test the effect on low- or high-income investors; the pattern emerges for all of them. Therefore, as a confirmation of the theoretical model, we find that previous outcomes indeed affect asset allocation and, subsequent to losses and gains, yield a money inflow into assets with higher and lower risk respectively. This finding is confirmed in existing literature on mutual fund activity as well, in which a negative relationship was found between returns and subsequent money inflows (Warther, 1995; Goetzman and Massa, 1999; Edelen and Warner, 1999) and between contemporaneous inflow of equity and bond funds (Goetzmann et al., 2000). Therefore, we argue that the model can capture and explain the unexpected changes in the demand for capital assets.

Estimating a volatility model

Based on the findings presented above, the empirical estimation of the theoretical model is presented in the followings. In this section the unit-root volatility model of OT is analysed. The α and β parameters are estimated for the return and volatility time series of the daily values of the CRSP value-weighted equity index using both weekly and monthly periods. The return is defined as the sum of the logarithmic daily returns. The volatility is calculated as the standard deviation of the returns during the given period; however, since this would show the daily volatility, it is multiplied by the ratio of the standard deviation of weekly returns divided by the standard deviation of daily returns (the adjustment to weekly from daily sampling). The estimation is based on simulating an error term of

$$e_t = r_t - (r_{f,t} + \beta\sigma_{t-1} + \alpha(r_{t-1} - r_{f,t-1} - \beta\sigma_{t-1})). \quad (7)$$

where $\alpha \in [-1,0]$ and $\beta \in [0,1]$. Here, the error term is not homoscedastic, therefore, we define the standardized error u_t as

$$u_t = \frac{e_t}{\sigma_t}. \quad (8)$$

Since these parameters are particularly sensitive to the underlying assumptions, first the distribution of the error is fitted based on maximum likelihood, where u_t is assumed to follow a scaled Student's-t distribution with $E(u_t) = 0$. Then, we apply a Kolmogorov-Smirnoff test to measure the significance of the difference between the empirical and estimated distribution functions. The higher the significance, the better the fit, therefore, the (α, β) pair yielding the highest p-value indicates the best fit of a distribution conditional to $E(u_t) = 0$; in other words, this pair is considered to provide the least significant error terms.

The numerical simulation results yield $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -0.03 \\ 0.21 \end{bmatrix}$ with a p-value of 0.85 and $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -0.07 \\ 0.32 \end{bmatrix}$ with a p-value of 0.93 using weekly and monthly periods respectively. Both results confirm that investors include previous outcomes as a negative proxy for their required return while the positive relationship between risk and required return stays intact. The particularly high p-values indicate that the error terms are fitted well using scaled Student's-t distributions; thus, the test results are robust.

It is worth mentioning that the model presented above describes the dynamics of the volatility of the whole market. However, as presented above, asymmetric volatility affects individual assets as well. We argue that this phenomenon stands on the fact that market and asset returns are highly correlated, especially in periods of greater continuous shocks (e.g. the financial crisis) that affect volatility significantly. So, we present a correlation test between the volatilities of the index and individual assets. The findings presented in Table 4 are consistent with the proposed reasoning for individual asymmetry.

Table 4: Correlation between market and asset volatilities

	Weekly analysis	Monthly analysis
Positive correlations	500	499
significant at 5%	497	492
Negative correlations	0	1
significant at 5%	0	0

According to weekly analysis, volatility correlation with the index is positive for all the 500 individual assets, although in three cases it is not significant. Nonetheless, these three latter assets (in particular, the equities with tickers "MNST", "NAVI" and "NWSA") have only become recently listed in the stock exchange, and therefore, correlation is tested on a much shorter interval than in the other cases. Hence, in these three cases the significance test yields low p-values due to the insufficient number of observations.

Applying monthly periods a similar pattern arises. Out of the 8 insignificant correlation coefficients 6 can be attributed to short available time series here as well. Altogether the positive correlation between individual assets is a robust pattern both in our weekly and monthly analysis, and hence, it is indeed a reasonable cause for the asymmetric volatility of individual assets.

CONCLUDING REMARKS

We find that the derivations of the theoretical model of Ormos and Timotity (2015) are empirically sound. Therefore, their recent theoretical explanation for asymmetric volatility is supported from both theoretical and empirical sides as follows.

First, we show that, in line with their findings, individuals tend to become less risk-averse (or risk-seeking until a given point) and more risk-averse subsequent to losses and gains respectively. This pattern confirms the existence of intertemporal mental framing, that is, investors tend to aggregate in time and adjust their portfolio accordingly.

Second, our empirical parameter estimation in discrete time indicates that the proposed model of OT indeed outperforms the simple random walk model: we confirm the significance of the predicted negative effect of previous outcomes on subsequent volatility, whereas, the positive relationship between simultaneous volatility and expected return remains significantly positive.

ACKNOWLEDGEMENTS

We gratefully acknowledge the help of Terry Odean, who provided us with the individual trading dataset. We also would like to express our gratitude for the thoughtful remarks of Adam Zawadowski, which have significantly contributed to our paper. We thank Zsolt Bihary and Niklas Wagner for their comments and suggestions of at the 6th Annual Financial Market Liquidity Conference 2015 at Corvinus University of Budapest. Mihály Ormos acknowledges the support by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Dusán Timotity acknowledges the support by the Pallas Athéné Domus Scientiae Foundation. This research was partially supported by Pallas Athene Domus Scientiae Foundation.

REFERENCES

- Barber, B.M., Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, 773-806.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307-327.
- Edelen, R.M., Warner, J.B. (1999). Why are mutual fund flow and market returns related? Evidence from high-frequency data. *Evidence from High-frequency Data*.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.
- Goetzmann, W.N., Massa, M. (1999). Index funds and stock market growth (No. w7033). National bureau of economic research.
- Goetzmann, W.N., Massa, M., Rouwenhorst, K.G. (2000). Behavioral factors in mutual fund flows. INSEAD.
- Lof, M. (2014). Rational Speculators, Contrarians, and Excess Volatility. *Management Science*.
- McQueen, G., Vorkink, K. (2004). Whence GARCH? A preference-based explanation for conditional volatility. *Review of Financial Studies*, 17(4), 915-949.
- Ormos, M., Timotity, D. (2015): In Search of Asymmetric GARCH models: A Loss-Aversion-Based Explanation of Heteroscedasticity, 6th Annual Financial Market Liquidity Conference, Budapest
- Thaler, R.H., Johnson, E.J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, 36(6), 643-660.
- Warther, V.A. (1995). Aggregate mutual fund flows and security returns. *Journal of financial economics*, 39(2), 209-235.

INTERTEMPORAL CHOICE AND DYNAMICS OF RISK AVERSION

Mihály Ormos
Dusán Timotity
Department of Finance,
Budapest University of Technology and Economics
Magyar tudosok krt. 2., 1117 Budapest, Hungary
ormos@finance.bme.hu and timotity@finance.bme.hu

KEYWORDS

Asymmetric volatility; Risk seeking; Prospect theory; TGARCH; EGARCH; Volatility dynamics; Market microstructure; Heuristic-driven trader

ABSTRACT

This paper provides a theoretical explanation for the heteroscedasticity of asset returns. In line with existing empirical results, our model yields an asymmetric relationship between stock return and volatility. Based on the simple assumptions that investors behave according to Prospect Theory and are subject to mental accounting in a dynamic setting, we analytically derive the unit-root versions of two of the best fitting heteroscedasticity models (EGARCH and TGARCH).

INTRODUCTION

Time-varying volatility (heteroscedasticity) of asset returns has attracted much research in the recent decades. Since the milestone papers of Engle (1982) and Bollerslev (1986) a great number of scholarly paper has been devoted to the topic. Their findings indicate that the phenomenon can be modeled by GARCH type models. Nevertheless, as the evidence shows below, no robust theoretical foundation has been proposed yet.

Another important aspect of the autoregression puzzle is the asymmetry in the volatility process. In particular, the phenomenon known as “asymmetric volatility” implies that changes in the price of the underlying asset are negatively correlated with the volatility of the subsequent period. Despite the wide amount of literature devoted to asymmetric volatility (Black, 1976; Christie, 1982; and Schwert, 1989), it still misses a robust explanation.

Until now three main explanations for this latter puzzle have been proposed. The first is the leverage effect noted by Black (1976), Christie (1982) and Schwert (1989). The authors assume that if the value of an equity drops, the firm becomes more leveraged, therefore, the volatility of equity returns rises according to the increased risk, hence, causing the negative relationship between return and subsequent volatility. They conclude, however, that volatility is indeed an increasing function of financial leverage, the effect by itself is not sufficient to account for the observed negative correlation.

The second explanation labeled as the volatility feedback hypothesis states that in cases of unexpected increase in volatility (e.g. exogenous shocks), expected volatility rises accordingly, and thus increasing the required return of the given asset in line with equilibrium asset pricing

models. This latter has an immediate negative impact on current stock price; therefore, it strengthens or weakens the magnitude of a previous shock subsequent to losses or gains respectively hence causing the asymmetry. Numerous papers on the topic provided evidence in support of both explanations (Pindyck, 1984; or Kim et al., 2004), yet recent studies still yield controversial results: on the one hand, Bollerslev et al. (2006) find that the analysis of high-frequency data indicates no significant volatility feedback, while on the other hand, Bekaert and Wu (2000) conclude that the leverage effect is insignificant.

The third main explanation is given by McQueen and Vorkink (2004) – their initial setting is the closest to ours – proposing that volatility autocorrelation is due to investors’ inclusion of the fluctuation of prices in their perceived utility (i.e. loss-averse behavior). The authors assume that volatility increases both following gains and losses as in volatility feedback hypothesis. This assumption comes from the paper of Barberis et al. (2001), which latter provides an asset pricing interpretation of Thaler and Johnson’s (1990) experiment of prospect theory in a dynamic setting. The Barberis-Huang-Santos (2001) (henceforth BHS) paper assumes that perception of losses (gains) is more (less) painful (delightful) when they are subsequent to prior losses (gains). In other words, this latter means that previous losses increase and previous gains decrease risk-aversion. However, BHS do not take into account the entire analysis of Thaler and Johnson; the authors do not focus on the finding that investors become risk-seeking following losses and risk-averse subsequent to gains if the opportunity of breaking-even is included in the choice set, which, in fact, almost always applies to asset returns. As we show in the following section, the individual dataset we use in this paper provides support for the latter hypothesis instead of the assumption of BHS. In other words, the pattern obtained in empirical tests suggest that, in contrast to BHS and McQueen and Vorkink, volatility indeed increases following losses and decreases subsequent to gains in order to allow or prevent breaking-even respectively.

It is also worth mentioning that the original setting in which autoregressive conditional heteroscedastic models were defined was the expected utility theory, that is, the dynamics of volatility (i.e. the standard deviation of asset returns in a given period) have been analyzed mostly in the setting of the mean-variance optimization of standard asset pricing models (e.g. the CAPM). However, contradictory results of this approach to utility perception have been well-documented, which would lead to biased

interpretations. Therefore, we build our theoretical model on an alternative approach: the prospect theory of Kahneman and Tversky (1979).

Based on Ormos and Timotity (2015), in our paper we apply this latter approach in a dynamic interpretation along with mental accounting and derive the microfoundations of heteroscedasticity. Our main findings are that (i) previous, unexpected shocks have negative, linear effect on the investors' required return; (ii) this pattern yields a negative effect of previous market shocks on market volatility; (iii) our setting provides the microfoundations of a unit-root, asymmetric, autoregressive volatility process similar to Threshold Generalized Autoregressive Conditional Heteroscedasticity (TGARCH) and Exponential Generalized Autoregressive Conditional Heteroskedastic (EGARCH) models in discrete and continuous time respectively.

The paper is structured as follows: in the subsequent section the theoretical setting is derived along with the main empirical findings of behavioral patterns necessary to interpret the theory. The last section summarizes the main conclusions of the paper and provides potential ways of further research.

THE MODEL

Previous market return has been shown to play a dominant role in the volatility dynamics of assets and is mainly responsible for the asymmetric response to shocks. We argue that this phenomenon can be explained by applying prospect theory in a dynamic setting.

Dynamics of the required return

Assuming that investors hold portfolios similar to the market portfolio, or at least that they diversify and hence invest into multiple assets, their portfolio is highly correlated with the market. In other words, a negative or positive market shock leads to losses and gains on investors' portfolios. Thaler and Johnson (1990) show in their experimental study that in such cases (if breaking even is in the choice set) investors become risk-seeking following losses and more risk-averse subsequent to gains; they aim to avoid realizing losses (exactly as in disposition effect (Shefrin and Statman, 1985 and Odean, 1998)) and are afraid of losing previous paper gains. This behavior comes from the S-shaped value function of loss-aversion: if we include the previous outcome as a reference point, the convexity of utility perception in the domain of losses results in risk-seeking behavior as the expected utility reaches its maximum at positive risk. In this specific case, considering the previous outcome as a fixed loss would cause greater pain than aggregating in time and hoping to break even; however, realizing the previous gain yields higher expected utility than taking the risk of losing the accumulated wealth.

Therefore, mental accounting (the mental aggregation or separation of pieces of information) in a dynamic setting leads investors to aggregate in time. Hence, they aim to obtain a given reference return at each period or, at least,

earn this return on average. That is, if we assume the rational expectations of outcomes as the reference point (Koszegi and Rabin, 2006), the subsequent required return is decreased by the previous abnormal return to be able to obtain the pre-defined reference return on average. Analytically the aforementioned is described with the following equation

$$\mu_t = \mu_{t-1} + \alpha(r_{t-1} - \mu_{t-1}) + r_{f,t} - r_{f,t-1}; \alpha \in [-1,0]. \quad (1)$$

where $r_{f,t}$, r_t and μ_t stand for the risk-free rate, the portfolio return and the required portfolio return of a given investor respectively. In particular, if we assume that investors form their expectations rationally and allocate their portfolio accordingly (i.e. they choose from the efficient portfolios), μ_t represents both the required and the expected return of their portfolio as the latter two become equal. Therefore, $r_{t-1} - \mu_{t-1}$ stands for the abnormal portfolio return in the previous period, which modifies the current period required (expected) return through α . The economic interpretation of this latter variable is defined as the sensitivity of an investor to mental accounting (the aggregation of previous outcomes). It would make no sense to assume that market participants adjust the required return by more than the previous shock itself; hence, we set its lower boundary at -1. Its negative value is due to the definition: aggregating in time increases or decreases the required return subsequent to losses or gains respectively. The $r_{f,t} - r_{f,t-1}$ term is added as the correction for the change in the risk-free rate or inflation.

It is worth mentioning here that autoregressive conditional heteroscedasticity (henceforth ARCH) models (Engle, 1982) were created in the setting of standard asset pricing models that are based on the expected utility theory (EUT); however, the latter would not yield the behavior described above. In contrast to prospect theory, EUT assumes concave utility in all domains of wealth, and therefore, would never induce risk-seeking behavior following losses. Therefore, the aforementioned behavior cannot be analyzed in a standard asset pricing structure, hence in order to give a coherent setting, the following section provides the definition of the risk-return relationship in prospect theory.

The mean-volatility relationship in prospect theory

The application of prospect theory in asset pricing attracted close attention in behavioral finance. Out of these, we discuss the most relevant findings that are related to our model. Levy and Levy (2004) argues that the mean-variance optimization of standard asset pricing models applies to prospect theory as well. In particular, they find that the prospect theory efficient set is a subset of the mean-variance efficient frontier and even by including probability distortion, the two sets almost coincide. Their results are confirmed and extended to asset pricing models in the paper of De Giorgi et al.

(2003) and Barberis and Huang (2008). The latter papers show that if the financial market equilibrium exists then the security market line theorem of CAPM holds under cumulative prospect theory as well. This finding also means that diversifying investors hold portfolios from the capital market line, and therefore, the relationship between volatility and expected return is linear for efficient portfolios.

Adding this linearity to the theory of the inclusion of previous gains and losses (as in Eq. (1)) leads to linearly decreased and increased portfolio volatility subsequent to gains and losses respectively.

The dynamics of portfolio volatility

In the followings, we present an analytical derivation of the dynamics of volatility. We define the intertemporal change of volatility in Eq. (2) and (3). Here we assume that in an equilibrium setting the price of risk does not change over time; nonetheless, the required return is not constant but follows the dynamics of

$$\mu_t = r_{f,t} + \beta\sigma_t = \mu_{t-1} + \alpha(r_{t-1} - \mu_{t-1}) + r_{f,t} - r_{f,t-1} = r_{f,t} + \beta\sigma_{t-1} + \alpha(r_{t-1} - r_{f,t-1} - \beta\sigma_{t-1}). \quad (2)$$

Here we applied the aforementioned linearity between risk and expected return of the CAPM setting. As long as we assume that investors hold well-diversified portfolios, only systematic risk is priced; therefore, σ_t stands for the market-related portfolio risk (henceforth volatility). β represents the slope of capital market line or the price of risk. The economic interpretation of Eq. (2) is that subsequent to losses investors aim to obtain higher expected return; however, according to equilibrium pricing, they can only achieve their goal by investing in riskier assets or increasing leverage. Solving the latter equation for the dynamics of volatility yields

$$\sigma_t = \sigma_{t-1} + \frac{\alpha}{\beta}(r_{t-1} - r_{f,t-1} - \beta\sigma_{t-1}) = \sigma_{t-1} + \frac{\alpha}{\beta}e_{t-1} = \sigma_{t-1} + \frac{\alpha}{\beta}\sigma_{t-1}\Delta W_{t-1} = \sigma_{t-1}\left(1 + \frac{\alpha}{\beta}\Delta W_{t-1}\right), \quad (3)$$

where e_{t-1} and ΔW_{t-1} represent a normally distributed error term and the change in the standard Wiener process in discrete time. Eq. (3) reveals that σ_t follows a unit-root process with constant conditional mean, that is

$$E[\sigma_{t+\tau}|\mathcal{F}_t] = \sigma_t + E\left[\sum_{i=t}^{t+\tau-1} \frac{\alpha}{\beta}\sigma_i\Delta W_i|\mathcal{F}_t\right] = \sigma_t + \frac{\alpha}{\beta}\sum_{i=t}^{t+\tau-1} E[\sigma_i|\mathcal{F}_t]E[\Delta W_i|\mathcal{F}_t] = \sigma_t + \frac{\alpha}{\beta}\Delta W_t, \quad (4)$$

where \mathcal{F}_t stands for the filtration (information available) at time t . Here, the separation of contemporaneous volatility and noise requires the assumption that they are uncorrelated (only the delayed response yields a negative correlation). According to Eq. (4), the volatility process

seems to be valid and realistic in the sense that periodical volatility tends to remain in a finite interval over a long horizon; it converges neither to infinity nor to zero.

Furthermore, Eq. (3) reveals another interesting pattern: it is very similar to the Threshold Generalized Autoregressive Conditional Heteroscedasticity (TGARCH) model introduced by Zakoian (1994) that is one of the most accurate heteroscedasticity models based on goodness-of-fit tests (Awartani and Corradi, 2005; Tavares et al., 2008). In particular, TGARCH models are defined as

$$\sigma_t = K + \delta\sigma_{t-1} + \alpha^+e_{t-1}^+ + \alpha^-e_{t-1}^- \quad (5)$$

where $e_{t-1}^+ = \begin{cases} e_{t-1} & \text{if } e_{t-1} > 0 \\ 0 & \text{if } e_{t-1} \leq 0 \end{cases}$ and $e_{t-1}^- = \begin{cases} e_{t-1} & \text{if } e_{t-1} \leq 0 \\ 0 & \text{if } e_{t-1} > 0 \end{cases}$. Therefore, the special case of Eq. (3) implies that $K=0$, $\delta=1$ and $\alpha^+ = \alpha^- = \frac{\alpha}{\beta}$. Effects of

previous gains and losses could be handled separately in Eq. (3) as well by using different α^+ and α^- ; however, as we show below, previous gains play only a much less significant role in the asymmetric effect on volatility. Nevertheless, distinct α^+ and α^- would also have a reasonable economic interpretation: considering that extreme gains do not cause a negative required return, that is, investors cannot and will not invest in assets with negative expected return irrespective of the previous outcomes, gains should have a less significant effect on the subsequently required return, therefore, α^+ should differ from α^- .

Another interpretation of Eq. (3) leads to another well-fitting, asymmetric GARCH model: the Exponential Generalized Autoregressive Conditional Heteroscedasticity (EGARCH) by Nelson, (1991). Dividing by σ_{t-1} and taking the natural logarithms of both sides yields

$$\ln \sigma_t = \ln \sigma_{t-1} + \ln \left(1 + \frac{\alpha}{\beta}\Delta W_{t-1}\right) \quad (6)$$

Taking the Taylor approximation around $\Delta W_{t-1} = 0$ then gives

$$\ln \sigma_t = \ln \sigma_{t-1} + \frac{\alpha}{\beta}\Delta W_{t-1} - \frac{1}{2}\left(\frac{\alpha}{\beta}\right)^2\Delta W_{t-1}^2 + \sum_{n=3}^{\infty} \frac{(-1)^{n-1}}{n!}\left(\frac{\alpha}{\beta}\right)^n\Delta W_{t-1}^n. \quad (7)$$

Due to the well-known property of the Wiener process, as Δt approaches to zero (the continuous time version is considered) third and higher order polynomials of ΔW_t vanish and $\Delta W_t^2 = \Delta t$. Therefore, the continuous time version of Eq. (7) can be written as

$$\ln \sigma_t = \ln \sigma_{t-1} + \frac{\alpha}{\beta} dW_{t-1} - \frac{1}{2} \left(\frac{\alpha}{\beta} \right)^2 dt, \quad (8)$$

or by multiplying both sides by 2

$$\ln \sigma_t^2 = \ln \sigma_{t-1}^2 + 2 \frac{\alpha}{\beta} dW_{t-1} - \left(\frac{\alpha}{\beta} \right)^2 dt. \quad (9)$$

The similarity to EGARCH comes from its definition of

$$\ln \sigma_t^2 = \omega + \beta_1 [\theta dW_{t-1} + \lambda (|dW_{t-1}| - E|dW_{t-1}|)] + \alpha_1 \ln \sigma_{t-1}^2, \quad (10)$$

where $\omega = - \left(\frac{\alpha}{\beta} \right)^2 dt$, $2 \frac{\alpha}{\beta} = \beta_1 \theta$, $\lambda = 0$ and $\alpha_1 = \alpha$ yields exactly Eq. (9). The unit-root, constant conditional mean property of Eq. (9) is again found by applying Itô's lemma for

$$x_t \equiv \ln \sigma_t^2, dx_t = 2 \frac{\alpha}{\beta} dW_t - \left(\frac{\alpha}{\beta} \right)^2 dt. \quad (11)$$

Then the inverse function is defined as

$$\sigma_t = e^{0.5x_t}. \quad (12)$$

By Itô's lemma

$$\begin{aligned} d\sigma_t &= \left[- \left(\frac{\alpha}{\beta} \right)^2 \frac{\partial \sigma_t}{\partial x_t} + \frac{1}{2} \left(2 \frac{\alpha}{\beta} \right)^2 \frac{\partial^2 \sigma_t}{\partial x_t^2} \right] + 2 \frac{\alpha}{\beta} \frac{\partial \sigma_t}{\partial x_t} dW_t = \\ &= \left[- \left(\frac{\alpha}{\beta} \right)^2 0.5\sigma_t + \frac{1}{2} \left(2 \frac{\alpha}{\beta} \right)^2 0.25\sigma_t \right] + \\ &+ 2 \frac{\alpha}{\beta} 0.5\sigma_t dW_{t-1} = \frac{\alpha}{\beta} \sigma_t dW_t. \end{aligned} \quad (13)$$

Again, the correlation between concurrent volatility and noise has zero expected value, therefore, the conditional mean is constant regardless of the length of delay. In conclusion, the TGARCH and EGARCH models are implications of prospect theory in a dynamic setting and they represent the underlying volatility process in discrete and continuous time respectively.

The dynamics of market volatility

We have derived so far the change of investors' risk attitude and the dynamics of the volatility of their portfolios. However, reasons behind the change of market volatility have not yet been covered. In this section, we propose an explanation for the positive relationship between the dynamics of market volatility

and the riskiness of investors' portfolio based on a simple market microstructural idea.

As discussed above, mental accounting leads to a clear pattern in investors choice that depends on the previous unexpected price shock: losses increase the subsequent demand for risky assets, whereas, gains reduce their demand. If we stick to the idea that investors hold the market portfolio or at least a well-diversified one that is correlated with the market, one can clearly see the following market microstructural situation: in line with the model of Glosten and Milgrom (1985) we find informed and uninformed traders in the market with probabilities π and $(1-\pi)$ that place market orders. In their model the informed investors know the exact value of an asset that can be either high (v^H) or low (v^L) and place their orders accordingly. Other participants of the market, such as the specialists that provide the liquidity by placing limit orders (thus define the spread) know only the probability of the true value that is $P(v = v^H) = \theta$ and $P(v = v^L) = 1-\theta$. Uninformed investors place their buy and sell orders completely randomly, hence, the probabilities of buy and sell orders coming from uninformed traders are equal ($P=0.5$).

Therefore, the profit of specialists is generated by the losses on transactions with informed investors and gains on transactions with uninformed investors. If we assume that the market is competitive, their zero expected profit criteria for transactions at the buy limit price and at the sell limit price yield the equilibrium ask and bid prices respectively (and the spread as their difference).

Moreover, if we introduce the pattern discussed in the previous sections, the spread changes in the following way: let us assume that, based on mental accounting heuristic, there is a new type of investors in addition to informed and uninformed traders, the heuristic-driven investor. This latter definition is not new in related literature: although, according to the pioneering papers of Glosten and Milgrom (1985) and Kyle (1985) uninformed traders are defined as those who do not possess fundamental information on assets, irrespective of their motives, a definition similar to our setting has already appeared in the paper of Bloomfield et al. (2009b), in which uninformed investors can have other trading motives than fundamental (e.g. behavioral). In their study the similar three-class distinction of investors is analyzed, where informed and uninformed investors and liquidity traders are present. The liquidity trader, however, may follow a behavioral pattern according to the dynamics of liquidity demand we have discussed so far, hence, we call this class the heuristic-driven trader.

Turning back to the introduction of such traders in the equilibrium criteria, let π and δ and $(1-\pi-\delta)$ stand for the shares of informed, heuristic-driven and uninformed traders (the probability of their trades). Then, subsequent to a negative market shock, the zero profit criteria of specialists at the ask and bid prices can be defined as

$$\pi(a - v^H) + \delta(a - v) + 0.5(1 - \pi - \delta)(a - v) = 0, \quad (14)$$

$$(1 - \theta)\pi(v^L - b) + 0.5(1 - \pi - \delta)(v - b) = 0. \quad (15)$$

Then the ask price is given as

$$\frac{\theta\pi v^H + 0.5(1 - \pi + \delta)v}{\theta\pi + 0.5(1 - \pi + \delta)} = v + \frac{\theta\pi(v^H - v)}{\theta\pi + 0.5(1 - \pi + \delta)} = v + \frac{\theta\pi(1 - \theta)(v^H - v^L)}{\theta\pi + 0.5(1 - \pi + \delta)}, \quad (16)$$

whereas the bid price follows

$$\frac{(1 - \theta)\pi v^L + 0.5(1 - \pi - \delta)v}{(1 - \theta)\pi + 0.5(1 - \pi - \delta)} = v + \frac{(1 - \theta)\pi(v^L - v)}{(1 - \theta)\pi + 0.5(1 - \pi - \delta)} = v - \frac{\theta\pi(1 - \theta)(v^H - v^L)}{(1 - \theta)\pi + 0.5(1 - \pi - \delta)} \quad (17)$$

One can clearly see the economic processes underlying in the aforementioned formulas: if herustic-driven traders are present the midprice differs from the expected value. Subsequent to a negative shock, the δ proportion of investors place buy orders at the ask price; however, they do not form supply at the bid price. Furthermore, their uninformed trades contribute positively to the profit; therefore, the equilibrium ask price declines as in Eq. (16). Still, their existence lowers the proportion of uninformed investors; hence, the equilibrium bid price declines as well as in Eq. (17). Although, both the ask and bid prices decline, the zero profit remains intact due to the modified probabilities of incoming buy and sell orders.

Then, the spread in competitive equilibrium can be defined as

$$S_- = \frac{\theta\pi(1 - \theta)(v^H - v^L)}{\theta\pi + 0.5(1 - \pi + \delta)} + \frac{\theta\pi(1 - \theta)(v^H - v^L)}{(1 - \theta)\pi + 0.5(1 - \pi - \delta)} = \frac{\theta\pi(1 - \theta)(v^H - v^L)}{[\theta\pi + 0.5(1 - \pi + \delta)][(1 - \theta)\pi + 0.5(1 - \pi - \delta)]} \quad (18)$$

where S_- stands for the spread subsequent to a negative market shock. The spread following positive market shocks is similar except for the sign of δ :

$$S_+ = \frac{\theta\pi(1 - \theta)(v^H - v^L)}{[\theta\pi + 0.5(1 - \pi - \delta)][(1 - \theta)\pi + 0.5(1 - \pi + \delta)]}. \quad (19)$$

Let the spread be defined as a function of Δ where

$$\Delta = \begin{cases} \delta & \text{for negative previous shocks} \\ -\delta & \text{for positive previous shock} \end{cases},$$

$$S(\Delta) = \frac{\theta\pi(1 - \theta)(v^H - v^L)}{[\theta\pi + 0.5(1 - \pi + \Delta)][(1 - \theta)\pi + 0.5(1 - \pi - \Delta)]}. \quad (20)$$

Then $S_- > S_+$ if and only if $S(|\Delta|) > S(-|\Delta|)$. As the numerator takes on a constant value in the function, we focus on the denominator value $f(\Delta)$. Then, $S_- > S_+$ if and only if $f(|\Delta|) < f(-|\Delta|)$, where

$$f(\Delta) = [\theta\pi + 0.5(1 - \pi + \Delta)][(1 - \theta)\pi + 0.5(1 - \pi - \Delta)] \quad (21)$$

is a concave, second order polynomial function of Δ . If and only if the maximum place of this function is reached in its negative domain, then $f(|\Delta|) < f(-|\Delta|)$ is always true. Therefore, it is enough to test whether

$$\operatorname{argmax}_{\Delta} f(\Delta) < 0. \quad (22)$$

According to the first order condition

$$0.5[(1 - \theta)\pi + 0.5(1 - \pi - \Delta)] - 0.5[\theta\pi + 0.5(1 - \pi + \Delta)] = 0, \quad \Delta = (1 - 2\theta)\pi. \quad (24)$$

Hence, if and only if $\theta > 0.5$, then $\operatorname{argmax}_{\Delta} f(\Delta) < 0$, $f(|\Delta|) < f(-|\Delta|)$, $S(|\Delta|) > S(-|\Delta|)$ and $S_- > S_+$. In other words, if the probability of a subsequent higher value is greater than that of a lower value, then spread is greater subsequent to a negative shock than it is following a positive shock.

The economic intuition behind an average $\theta > 0.5$ is simple as the growth of value is one of the basic assumptions in analyzing capital markets. This greater probability of a higher value is confirmed by empirical studies as well: although the authors apply a bit different methodology, Easley et al. (2002) and Brennan et al. (2014) measure the probability of an increase in the value to be $P(v|v = v^H) = 0.67$ and 0.614 .

In conclusion, we argue that, on average, the spread increases subsequent to losses and decreases subsequent to gains. Moreover, considering that continuous market orders at the ask and bid prices define the standard deviation of price changes, our explanation clearly implies that previous positive (negative) shocks decrease (increase) both the spread and the volatility accordingly. Related literature provides further support to our aforementioned reasoning. Park and Sabourian (2011) analyze a similar setting based on the Glosten-Milgrom model and find that people act as contrarian if their information leads them to concentrate on middle values. Kaniel et al. (2008), Choe et al. (1999), Grinblatt and Keloharju (2000, 2001), Richards (2005), Bloomfield et al. (2009a) also confirm the existence of such contrarian traders. Moreover, according to Lof (2014), the introduction of contrarian trading in asset pricing models dramatically increases the predictive power of the models. Furthermore, our former, mental accounting-based explanation for the contrarian activity is supported

by Yao and Li (2013) who argue that prospect theory investors can behave as contrarian noise traders in a market, while Kadous et al. (2014) finds that investors act as contrarians if and only if they have held in the past the particular asset that they buy in the subsequent period; this latter provides evidence that mental accounting and prospect theory are indeed responsible for the negative feedback trading instead of an alternative exogenous factor. For the well-documented, significant, positive relationship between spread and price volatility see Hussain (2011) Wang and Yau (2000), Wyart et al. (2008).

CONCLUDING REMARKS

We find that asymmetric and autoregressive volatility measured in previous empirical studies in asset pricing can be derived from and attributed to intertemporal choice of investors, assuming that they behave according to prospect theory in a dynamic setting. We show that, in contrast to most of the studies on this topic, individuals should tend to become less risk-averse (or risk-seeking until a given point) and more risk-averse subsequent to losses and gains respectively, which leads to the rejection of the volatility feedback and BHS explanations for asymmetric volatility. Furthermore, we argue that the third existing explanation (the leverage effect) does not hold either, as we find a volatility decreasing effect of both previous gains and losses of a given asset when controlling for the market return.

However, our proposed model is based on a negative relationship between market returns and market volatility; and is thus able to capture the dynamics of volatility measured empirically. Combining the linear relationship between risk and return, as presented above in detail, and the aforementioned pattern in the intertemporal choice (i.e. the required return) yields the autoregressive conditional heteroscedasticity model presented in this paper. We show that the discrete and continuous time alternatives of the main equation result in the TGARCH and EGARCH models respectively, which in particular are measured to be two of the regressions with the highest goodness-of-fit in most of the empirical studies.

Potential ways of further research include various opportunities. First, an experimental analysis would be interesting to show whether these patterns are found in a laboratory environment as well if the focus is on the effect of breaking-even. Second, the influence of this behavior on asset liquidity and market microstructure could be analyzed in detail including an empirical analysis of the probability estimation of heuristic-driven traders. Third, the application of the proposed model in mathematical finance could reveal further interesting patterns; in particular, asymmetric stochastic volatility models (Heston and Nandi, 2000) in option pricing are found to provide better estimates on option prices and fit the “volatility smile” of the Black-Scholes implied volatilities, which regressions could be further improved by including the proposed model described in this paper. Finally, the introduction of cognitive research, such as

the neuroeconomic approach, could reveal further underlying factors behind the behavioral patterns presented in this paper.

ACKNOWLEDGEMENTS

We gratefully acknowledge the help of Terry Odean, who provided us the individual trading dataset. We also would like to express our gratitude for the thoughtful remarks of Adam Zawadowski, which have significantly contributed to our paper. We thank Zsolt Bihary and Niklas Wagner for their comments and suggestions of at the 6th Annual Financial Market Liquidity Conference 2015 at Corvinus University of Budapest. Mihály Ormos acknowledges the support by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Dusán Timótiy acknowledges the support by the Pallas Athéné Domus Scientiae Foundation. This research was partially supported by Pallas Athene Domus Scientiae Foundation.

REFERENCES

- Awartani, B.M., Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting*, 21(1), 167-183.
- Barberis, N., Huang, M., Santos, T. (2001). Prospect Theory and Asset Prices. *Quarterly Journal of Economics*, 116, 1-53
- Barberis, N., Huang, M. (2008). Stocks as Lotteries: The Implications of Probability Weighting for Security Prices. *American Economic Review*, 98, 2066-2100.
- Bekaert, G., Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies* 13, 1-42.
- Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economical Statistics Section* 177-181.
- Bloomfield, R., O'hara, M., Saar, G. (2009b). How noise trading affects markets: An experimental analysis. *Review of Financial Studies*, 22(6), 2275-2302.
- Bloomfield, R. J., Tayler, W. B., Zhou, F. H. (2009a). Momentum, reversal, and uninformed traders in laboratory markets. *The Journal of Finance*, 64(6), 2535-2558.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307-327.
- Bollerslev, T., Litvinova, J., Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, 4(3), 353-384.
- Choe, H., Kho, B. C., Stulz, R. M. (1999). Do foreign investors destabilize stock markets? The Korean experience in 1997. *Journal of Financial Economics*, 54(2), 227-264.
- Christie, A.A. (1982). The stochastic behavior of common stock variances – value, leverage and interest rate effects. *Journal of Financial Economics* 10, 407-432.
- Easley, D., Hvidkjaer, S., O'hara, M. (2002). Is information risk a determinant of asset returns?. *The journal of finance*, 57(5), 2185-2221.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 987-1007.

- Glosten, L.R., Milgrom, P.R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics*, 14(1), 71-100.
- Grinblatt, M., Keloharju, M. (2000). The investment behavior and performance of various investor types: a study of Finland's unique data set. *Journal of financial economics*, 55(1), 43-67.
- Grinblatt, M., Keloharju, M. (2001). How distance, language, and culture influence stockholdings and trades. *Journal of Finance*, 1053-1073.
- Heston, S. L., Nandi, S. (2000). A closed-form GARCH option valuation model. *Review of Financial Studies*, 13(3), 585-625.
- Hussain, S. M. (2011). The intraday behaviour of bid-ask spreads, trading volume and return volatility: evidence from DAX30. *International Journal of Economics and Finance*, 3(1), 23-34.
- Kadous, K., Tayler, W. B., Thayer, J. M., Young, D. (2014). Individual Characteristics and the Disposition Effect: The Opposing Effects of Confidence and Self-Regard. *Journal of Behavioral Finance*, 15(3), 235-250.
- Kaniel, R., Saar, G., Titman, S. (2008). Individual investor trading and stock returns. *The Journal of Finance*, 63(1), 273-310.
- Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.
- Kim, C.J., Morley, J., Nelson, C. (2004). Is there a Positive Relationship between Stock Market Volatility and the Equity Premium? *Journal of Money, Credit, and Banking* 36, 339-360.
- Kószegi, B., Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 1133-1165.
- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 1315-1335.
- Levy, H., Levy, M. (2004). Prospect theory and mean-variance analysis. *Review of Financial Studies*, 17(4), 1015-1041.
- Lof, M. (2014). Rational Speculators, Contrarians, and Excess Volatility. *Management Science*.
- McQueen, G., Vorkink, K. (2004). Whence GARCH? A preference-based explanation for conditional volatility. *Review of Financial Studies*, 17(4), 915-949.
- Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance*, 53(5), 1775-1798.
- Ormos, M., Timotity, D. (2015): In Search of Asymmetric GARCH models: A Loss-Aversion-Based Explanation of Heteroscedasticity, 6th Annual Financial Market Liquidity Conference, Budapest
- Park, A., Sabourian, H. (2011). Herding and contrarian behavior in financial markets. *Econometrica*, 79(4), 973-1026.
- Pindyck, R.S. (1984) Risk, inflation, and the stock market. *American Economic Review* 74, 334– 351.
- Richards, A. (2005). Big fish in small ponds: The trading behavior and price impact of foreign investors in Asian emerging equity markets. *Journal of Financial and quantitative Analysis*, 40(01), 1-27.
- Shefrin, H., Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of finance*, 40(3), 777-790.
- Schwert, G.W. (1989). Why does stock market volatility change over time? *Journal of Finance* 44, 1115–1153.
- Tavares, A.B., Curto, J.D., Tavares, G.N. (2008). Modelling heavy tails and asymmetry using ARCH-type models with stable Paretian distributions. *Nonlinear Dynamics*, 51(1-2), 231-243.
- Thaler, R.H., Johnson, E.J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, 36(6), 643-660.
- Wang, G. H., Yau, J. (2000). Trading volume, bid-ask spread, and price volatility in futures markets. *Journal of Futures markets*, 20(10), 943-970.
- Wyart, M., Bouchaud, J. P., Kockelkoren, J., Potters, M., Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1), 41-57.
- Zakoian, J.M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, 18(5), 931-955.
- Yao, J., Li, D. (2013). Prospect theory and trading patterns. *Journal of Banking & Finance*, 37(8), 2793-2805.

ESTIMATION OF CUSTOMER DEFAULT BASED ON BEHAVIOURAL VARIABLES

Nóra Felföldi-Szűcs
Department Finance
Kecskemét College, Corvinus University of Budapest
Izsáki street 10, Kecskemét 6000, Hungary
E-mail:szucs.nora@gamf.kefo.hu

KEYWORDS

Scoring models, partner risk, credit risk.

ABSTRACT

The paper focuses on the estimation of customer default among the small and medium enterprises (SME). Based on the literature on credit scoring models we build a logistic regression model which is widely used by commercial banks. Our models predicting customers' default on their payables to suppliers are estimated on a sample of a customer portfolio of 905 SME clients. Based on the analysis the non-financial, behavioural variables estimate better customer default than the financial ratios. Our models perform weaker than the usual performance level of scoring models in commercial bank. This result assumes that defaulting on a payable to suppliers is an early signal of possible financial difficulties.

INTRODUCTION

As the actors in all kind of credit contracts even firms offering trade credit to their customers are exposed to credit risk. This idea is built into their practice when making decisions on offering delayed paying called trade credit to the group of reliable customers but requiring prompt payment from less opac partners. The decision made by the firms is based on all the important information considered by commercial banks when offering credit to their borrowers. Hago (Hago 2001) describes in his paper the corporate credit policy and as part of it the corporate credit analysis.

Based on the less formal practice of a huge part of firms we will formalize our analysis and we will apply the credit risk modeling methodology of banks to the customer portfolio of a firm trading in construction materials. We compare the classification accuracy of financial and behavioural variables in the case of SME customers. The findings harmonize with the practice of the claim management firm who provided me the database. The paper describes the applied methodology and the database. After forming the hypothesis we estimate the models forecasting customers' nonpayment. After the results we finally conclude.

THE APPLIED METHODOLOGY AND THE DATASET

The literature of credit and default risk modeling is rather abundant, and what is more, these keywords often lead to writings with surprisingly differing contents. From a historical approach, it is the accounting-based, so-called credit risk scoring models we will first encounter in literature. Accounting-based models are based on financial ratios derived from the financial / accounting statements of the companies; according to the values taken by these ratios, businesses are divided into two groups: bankrupt and solvent firms. More on accounting based models can be found in the works of a number of international and Hungarian authors, like (Altman and Saunders 1997, Liao et al. 2005, Platt and Platt 1990 or Kiss 2003, Virág 2004 and Oravecz 2008. Relevant pieces of literature clearly distinguish between loans for SMEs and those for large corporations, thus the related risk assessment methodologies are reasonably expected to be different, too. Authors focusing directly on SME borrowers mention the logistic regression as the most widespread model (Atiya 2001 és Laitinen and Laitinen 2000) and most of the authors uses logistic regression for their own estimates of firms' default (Altman and Sabato 2007, Falkenstein et al. 2000). We will follow their practice in this research.

The trade credit database consists of the customer portfolio of a real-life company. This business is a member of a multinational group of corporations with several subsidiaries in Hungary, trading in construction materials. The total open receivables are of 2.6 billion HUF (ca. 8 million EUR), the delayed receivables of 1.4 billion HUF equal the sales revenue for 46 days. Besides the open receivables totals from all the 905 SME customers of the company, a record of overdue amounts and an aged balance of accounts receivable was also provided. These being stock variables, the figures relate to one specific day. In addition to the agreed credit limit, information (partly of a qualitative nature) on the customer, its manager and its payment history also appear in the database; these will be included in the quantitative analysis as dummy variables. Thus the variables that are given or can be defined for each and every customer are as follows: Aged balanced of open and overdue receivables; detailed breakdown of open and overdue receivables by due date as of the date

examined; the amount (if any) purchased/paid back between the two dates can be established; how many times the customer appeared on the so-called blacklist (record of non-paying customers) of the claims management company; whether the owner/manager has held a similar position in a company that went bankrupt or had to be liquidated; whether there is anything suspicious about the company (tax (and similar) arrears, foreclosure initiated against the company, frequent changes in place of residence and scope of activities, the credit line extended by the supplier, if any, the amount (if any) by which the credit line was exceeded) can be established.

Non-payment was defined through a dummy variable (DEF90) which equals 1 if the customer is more than 90 days past due, 0 otherwise. An important remark to the above is that these definitions do not coincide with the criteria of bankruptcy and even less so with those of the company's liquidation – they intend to describe a less extreme situation when non-payment „only” affects the supplier. Variable DEF90 is primarily based on the New Basel Capital Accord (Basel II), which defines a defaulted borrower as anyone who is more than 90 days behind with their payments (BIS 2006). Even though our own definition of DEF90 and that of Basel II takes the exact same form, an important distinction is to be made depending on whom the client is indebted to. We made the assumption that it is companies' suppliers who first suffer from late payments, and it is only afterwards, if further financial difficulties arise, that they dare fall behind with or default on their obligations to banks. Accordingly, our nonpayment variables describe a situation 'weaker' than either bankruptcy or a default on a bank loan, which must be taken into account when constructing our model and when interpreting the findings. But variable DEF90 defines a delay of payment far more larger than the average days of delay of the portfolio thus one can assume that it captures the difference on a delay and of a default. This weighted average delay of the customers are 55 days.

As a final step in data collection, we also looked up the company's key balance sheet and income statement figures in order to aid our later analyses. The financial ratios used during modeling are as follows:

- Total Liabilities/(Total Liabilities + Owner's Equity)
- Earnings Before Taxes/Net Sales Revenue
- Earnings Before Taxes/Total Assets
- EBIT/Total Assets
- EBITDA/Net Sales Revenue
- EBIT/Net Sales Revenue
- Net Earnings/Owner's Equity (ROE)
- Current Assets/Current Liabilities
- Total Liabilities/(EBIT + Income from Financial Transactions)
- Total Liabilities/EBITDA

- EBIT/Expenses on Financial Transactions
- Current Liabilities/Net Sales Revenue
- Current Assets/Total Assets
- Total Receivables/Total Liabilities
- Owner's Equity/Fixed Assets
- Net Sales Revenue/Total Assets
- Net Sales Revenue/Net Working Capital
- Net Sales Revenue/EBIT
- (Earnings Before Taxes+Expenses on Financial Transactions)/Total Assets
- Profit on Ordinary Activities/Owner's Equity
- Net Working Capital/Total Assets
- Cash and Cash Equivalents/Current Liabilities
- Long-Term Debt/Owner's Equity
- Total Receivables /Owner's Equity
- Long-Term Debt/ (Total Liabilities + Owner's Equity)
- Total Receivables/(Total Liabilities + Owner's Equity)
- Net Sales Revenue/Net Working Capital
- Cash and Cash Equivalents/Total Assets
- Current Liabilities/Owner's Equity
- Cash and Cash Equivalents/Net Sales Revenue
- (Net Sales Revenue $t=1$ /Net Sales Revenue $t=0$) -1
- FCFE/Total Assets

As many others had used it in bankruptcy modeling, we also used logistic regression to predict non-payment; from amongst the simpler methods, this is the most widely used one and it is considered rather successful, as well (Falkenstein 2000; Grunert et al 2005). Relying on relevant literature (Altman and Sabato 2007; Falkenstein 2000; in Hungary Kristóf 2008a-b) each model variation employed the Forward Stepwise Likelihood Ratio algorithm with significance levels of 5 percent and 10 percent for entry and removal, respectively. The sample was partitioned into a training and a holdout sample according to the 75% - 25% ratio recommended by literature (e.g. Imre 2008).

The studies we read all determined the cutoff value in very different ways. The cutoff value of the model is a threshold for the estimated probability of default: if the latter is lower /higher than the cutoff value then the model predicts the client in question to pay on time /to default on the payment, respectively. Oravecz (2008) and Tang-Chi (2005) discuss the determination of cutoff values for default prediction models in detail. Oravecz (2008) distinguishes between theoretical and empirical determination. The theoretical method relies on profit matrices. Money should be lent to the client as long as the expected profit of lending is higher than the expected profit of refusal. Oravecz (2008) even provides a numerical example and according to her empirical results, the cutoff should rather be determined using the theoretical method if and when profit maximization is the goal. Empirical approaches examine the model's

effectiveness for different cutoff values. Yet each author has their own interpretation of effectiveness. Oravec (2008) sticks with profit maximization, while Tang-Chi (2005) offer a number of different solutions. They cite Altman (1968) having chosen cutoffs based on classification accuracy. Frydman, Altman and Kao (1985), for example, minimized the number of misclassifications, while Ohlson (1980) opted for the intersection of the probability distributions of good and bad debtors.

Current literature primarily features cutoffs given by the largest AUC (area under the curve), arrived at by comparing AUC values calculated using a number of different cutoff values and choosing the one generating the maximum AUC. This is also the method we are going to use in our paper.

THE HYPOTHESIS

Our hypothesis is that the classification accuracy of the models relying solely on non-financial variables is not worse than that of the models using financial data only. Even though the range of non-financial information available to me is rather limited, we still intend to compare the discriminative power of financial statement data with that of other, non-financial data based on Altman, Sabato and Wilson (2010) and Lehman (2003). One of the motives for formulating this hypothesis was that the claims management company that provided me with the database had made recommendations to its client – the supplier – on the line of credit to be extended to each customer primarily based on non-financial indicators, that is, on a kind of expert system.

MODELS PREDICTING DEFAULT ON CUSTOMER RECEIVABLES

The model variation named „MULTIVAR_FIN_015” uses nothing else but publicly available financial data (financial ratios and publicly available blacklists of financially distressed firms, but no behavioural indicators), thus it can be used for new customers, too. The number “015” indicates that the optimal (AUC-maximizing) cutoff value is 15 percent. Accordingly, clients are classified as good debtors if their estimated probability of default is below 15 percent, and “bad” (i.e. non-paying) customers otherwise. For this very model, the results are presented in detail. For the second model only a shorter overview will be available.

Table 1: Parameters of model MULTIVAR_FIN_015

Source: SPSS

Name of variables	B	S.E.	Wald	Df	Sig.	Exp(B)
number of blacklist appearances	.245	.087	8.023	1	.005	1.278
Total liabilities/Total Debts	2.46	.404	36.274	1	.000	11.429

Owners' Equity/Fixed Assets	.005	.002	3.732	1	.053	1.005	
Net Sales Revenue/Total Assets	-	.086	6.882	1	.009	.798	
Cash and Cash Equivalents/Total Assets	1.786	.674	7.026	1	.008	5.964	
FCFF/Total Assets	.775	.209	13.734	1	.000	2.171	
Constant	-	.183	.347	84.241	1	.000	.041

According to the SPSS-output, the significant explanatory variables of customer default in the case of new customers are: the number of blacklist mentions, Total Liabilities/Total Debt, Net Sales Revenue/Total Assets, Cash and Cash Equivalents/Total Assets, and FCFF/Total Assets. The fact, for example, that Customer ‘A’ has been mentioned on a blacklist one single time results in their odds becoming 1.278 times the odds of an arbitrary Customer ‘B’ whose significant variables are identical to those of Customer ‘A’ except that Customer ‘B’ has never been added to any blacklist.

Table 2: Goodness-of-fit indices for model MULTIVAR_FIN_015

Source: SPSS

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerk R Square
1	511,963	0,058	0,099
2	498,222	0,079	0,134
3	490,464	0,09	0,154
4	483,7	0,1	0,17
5	476,435	0,11	0,188
6	470,034	0,119	0,204

From amongst the goodness-of-fit indices, Nagelkerke R² is the easiest to interpret, because it works like the coefficient of multiple determination, taking values between 0 and 1 (Oravec 2008). Consequently, the explanatory power of our model relying solely on publicly available financial information, is 20.4 percent.

As a next step we estimated a model based on behavioural variables (MULTIVAR_BEHAV_015). Even though the studies discussed in the methodological chapter used a rather wide range of data, our database was limited to the following variables: legal form of the company, repayment, number and duration of blacklist appearances, track record of the company and related persons, and the existence and the exceeding of a credit line. Therefore this model, similar to Altman’s ZETA-

model, also includes the $\ln(\text{Total Assets})$ indicator as a proxy variable of company size. Similarly, negative Owner's Equity balances were also taken into account through a dummy variable. Final results are listed in Table 3. The indicators found to be significant were: track record of the company (*comphist_dummy*), payment habits, exceeding of the credit line and negative owner's equity.

Table 3: Parameters of model MULTIVAR_BEHAV_015
Source: SPSS

Name of variables	B	S.E.	Wald	Df	Sig.	Exp(B)
number of blacklist appearances	0,264	0,102	6,664	1	0,01	1,303
number of blacklist days	0,004	0,002	3,725	1	0,05	1,004
firmhistory_dummy	-0,614	0,271	5,156	1	0,02	0,541
repayment_dummy			6,552	2	0,04	
repayment_dummy(1)	-0,4	0,268	2,22	1	0,14	0,67
repayment_dummy2	-0,968	0,384	6,354	1	0,01	0,38
exceeding creditline_dummy	-1,528	0,247	38,305	1	0	0,217
negative equity_dummy	1,562	0,414	14,233	1	0	4,767
Constant	-0,258	0,307	0,707	1	0,4	0,772

Table 4: Goodness-of-fit indices for model MULTIVAR_BEHAV_015
Source: SPSS

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	502,803	0,072	0,123
2	486,494	0,096	0,164
3	473,59	0,114	0,196

4	468,116	0,122	0,209
5	460,938	0,132	0,226
6	457,414	0,137	0,234

RESULTS

Based on the literature on relevant methodologies, we examined the hypothesis concerning the logit models classifying customers either as payers or non-payers. The comparison of our models also serves the purpose of evaluating the hypothesis. The aspects of comparison are listed in Tables 3 and 4 showing three goodness-of-fit indices. The estimation algorithm minimizes the value of -2Loglikelihood , thus: the lower the better. Concerning Cox-Snell R^2 values, however, it is the higher values that are more favorable. This indicator, by the way, compares the likelihood value to the empty model (Oravecz, 2008, Sajtos and Mitev, 2007). The interpretation of Nagelkerke R^2 has already been discussed earlier.

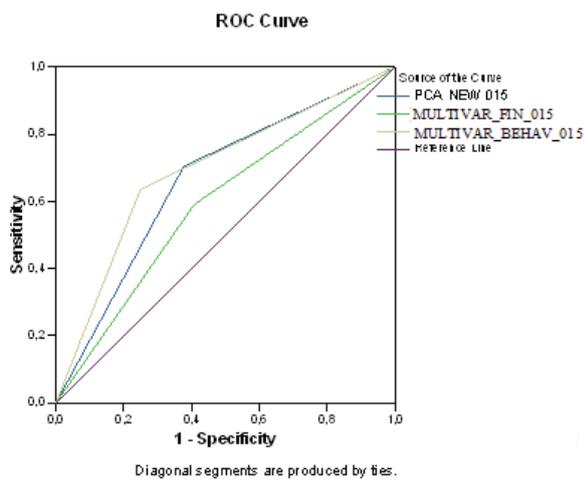
Model MULTIVAR_BEHAV_015, only employing behavioural and non-financial indicators as explanatory variables, was estimated for the purpose of testing this hypothesis. The goodness-of-fit indices and the AUC values of both the training sample and the holdout sample (see Tables 5 and 6) all support that replacing financial ratios with variables describing other dimensions of companies' behavior yielded a better-performing model. Based on the presented models, hypothesis has been accepted, that is, the classification accuracy of the models relying solely on behavioural variables is not worse than that of the models using financial data only. As an interesting note: the acceptance of hypothesis also explains the practice of the claims management company providing our database – namely, that they can successfully determine the credit lines to be extended to customers based primarily on behavioural variables and only secondarily on financial data.

Table 5: Testing of hypothesis – goodness-of-fit indices
Source: SPSS

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Entire sample			
MULTIVAR_FIN_015	470,034	0,119	0,204
MULTIVAR_BEHAV_015	457,414	0,137	0,234

Table 6: Testing of hypothesis – AUC
Source: SPSS

Training sample	AUC	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
MULTIVAR_FIN_015	0,686	0,029	0,000	0,628	0,743
MULTIVAR_BEHAV_015	0,703	0,029	0,000	0,646	0,760
Test sample					
MULTIVAR_FIN_015	0,591	0,048	0,063	0,497	0,686
MULTIVAR_BEHAV_015	0,693	0,047	0,000	0,602	0,785



Figures 1: ROC curves for the holdout sample
Source: SPSS

CONCLUSIONS

The success of the MULTIVAR_BEHAV_015 model suggests that the receivables managing company could improve its decision making mechanism by collecting more behavioural information. The literature recommends for instance the age of the customer relationship, the age of the buying company, the number of the employees, the education of the leaders of the firm, the leader's experience measured in years in the industry, the variability of the balance of the received trade credit, the industry and its industrial bankruptcy rate.

There is also a further research question related, namely to examine the classification power of other nonfinancial

indicators. The goodness of fit and the classification power of the models are slightly weaker than the similar values of the bankruptcy and scoring models. A possible reason is that suppliers are generally paid late. A delay on supplier payables does not mean such a severe event of credit risk with serious consequences like bankruptcy or a delay towards a bank. Imre (2008), who developed models for the prediction of bank loan defaults (delays beyond 90 days, in accordance with the default-definition of Basel II), drew the same conclusion at the end of his dissertation. Thus, most probably, the financial data of bankrupt businesses can be better distinguished from that of non-bankrupt businesses than the data of payers can be from that of non-payers. Adopting the reasoning of Imre (2008), a delay beyond 90 days on one's bank loan payment is a "weaker" event than bankruptcy, yet it is an even less severe credit risk situation if it is „only" the supplier who has to wait more than 90 days for their money. So late payment to suppliers is such an early signal for possible financial difficulties that the financial data of the firm can not reflect yet. Consequently, we regard the goodness-of-fit indices and the AUC values of our models as appropriate in spite of the fact that literature frequently reports of better performing models.

All of this however brings up another research question: could the models be improved if we reformulated the definition of customer non-payment which was the dependent variable in the logit models? This non-payment definition would be probably customized to the industry which the customer belongs to. It has to be an early signal about illiquidity and insolvency to assure that the supplier has still enough time to make suitable steps for the collection of the receivables. On the other hand the delay classified as non-payment should be sufficiently long to differ from the common, average delays of 50-60 days in the examined portfolio, so it can be modeled as a dependant variable and can be predicted in advance.

There would be additional research possibilities for the future if chronological data would be available for the aged balance of open receivables. First the circle of the behavioural variables could be broadened by a detailed knowledge on historical paying and purchasing habits. Second, the stability of the paying patterns could be tested. There is an interesting question, whether a customer from the current database classified as a delayer between 31-60 days was in the same due date interval in an earlier point of time, or he/she had belonged to the group of 16-30 days delayers earlier. This last finding would mean that the client is permanently falling behind towards the longer delays. It is also possible, that until a particular due date interval the classification is stabile, afterwards the customer stops his/her payments and his/her classification is going to be worse by the time. If the latter supposition is true, then the observation of this threshold in the due date

structure can help to construct a non-payment definition. If the historical value of open balances is available, then there is an opportunity to control and to test the results and the prediction power of the logistic models which are classifying the paying and the non-paying customers.

REFERENCES

- Altman, E.I. and G. Sabato. 2007. "Modelling Credit Risk for SMEs: Evidence from the U.S. Market". *Abacus* Vol. 43. No. 3. . 332–357.
- Altman, E. I., G. Sabato and N. Wilson. 2010. "The Value of Non-Financial Information in SME Risk Management". *Journal of Credit Risk*, Vol. 6, No. 2, . 5-25.
- Atiya, A.F. 2001. "Bankruptcy prediction for credit risk using neural networks: A survey and new results". *IEEE Transactions on Neural Networks*, Vol. 12. No. 4. . 929-935.
- Falkenstein, E. G., A. Boral and L. Carty. 2000. "RiskCalc for Private Companies: Moody's Default Model". *Global Credit Research*, May 2000.
- Grunerta, J., L. Norden and M. Weber. 2005. "The role of non-financial factors in internal credit ratings". *Journal of Banking & Finance*, Vol. 29, No. 2. . 509-531.
- Hago, T. M. 2001. "Some problems of trade credit". *Budapest Management Review*, Vol. 32. No. 3. . 27-40.
- Imre, B. 2008: Predicting default defined by Basel II - models based on Hungarian firms' data between 2002 and 2006. *PhD. dissertation*, University of Miskolc.
- Kiss, F. 2003. The development and application of A credit scoring. *PhD. dissertation*, Budapesti University of Technology and Economics
- Kristóf, T. 2008a. "A On methodological questions of bankruptcy prediction and PD estimation". *Economic Review*, Vol. 55. No. 5. . 441-461.
- Kristóf, T. 2008b. "Forecasting survival and paying ability of economic organizations". *PhD. dissertation*. Corvinus University of Budapest
- Laitinen, E. K. and T. Laitinen. 2000. "Bankruptcy prediction Alication of the Taylor's expansion in logistic regression". *International Review of Financial Analysis*, Vol. 9. No. 4. . 327-349.
- Lehmann, B. 2003. "Is It Worth the While? The Relevance of Qualitative Information in Credit Rating" (April 17, 2003). *EFMA 2003 Helinski Meetings*. Available at SSRN: <http://ssrn.com/abstract=410186> or doi:10.2139/ssrn.410186
- Oravec, B. 2007. "Credit scoring models and their performance". *Credit Institutes' Review*, Vol. 6. No. 6. . 607-627.
- Oravec, B. 2008. "Selectional bias and its reductions by credit scoring models". *PhD. dissertation*, Corvinus University of Budapest
- Sajtos, L., and A. Mitev. 2007. "SPSS research handbook". *Alinea*, Budapest
- Tseng-Chung Tang and Li-Chiu Chi 2005. "Predicting multilateral trade credit risks: comparisons of Logit and Fuzzy Logic models using ROC curve analysis". *Expert Systems with Applications*, Vol. 28, No. 3, . 547-556.
- Virág, M. 2004. "History of default prediction models." *Budapest Management Review*, Vol. 35. No. 10. . 24-32.

AUTHOR BIOGRAPHY

NÓRA FELFÖLDI-SZÚCS was born in Zalaegerszeg, Hungary and went to the Corvinus University of Budapest, where she studied financial investment analysis and risk management and obtained her degree in 2006. After a short experience at Deutsche Bank she began her PhD studies at Corvinus University where she has been lecturer since 2006. She obtained her PhD degree in 2013. Since 2015 she has been the coordinator of Business Administration Program at Kecskemét College. Her e-mail address is : nora.szucs@uni-corvinus.hu.



FACTORS AFFECTING HOUSEHOLD PARTICIPATION IN SOLID WASTE MANAGEMENT SEGREGATION AND RECYCLING IN BANGKOK, THAILAND

Walailak Atthirawong
Department of Statistics, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang, Bangkok10520, Thailand
Email:walailaknoi@gmail.com

KEYWORDS

Household Participation, Multiple Regression Analysis, Recycling, One-way Analysis of Variance, Segregation, Solid Waste Management

ABSTRACT

The number of population in Bangkok, the capital of Thailand, is increasing every year. The capital produces about 9,900 tons of garbage daily or 1.53 kilograms per person which only 13% of waste is recycled per day. This presents a serious challenge and concern of municipal authority in solid waste management. This study examines Bangkok residents' practices, knowledge of waste management, as well as the level of community mobilization and the level of household participation in solid waste segregation and recycling. One-way Analysis of Variance (ANOVA) was employed to test whether there was statistically significance between the level of household participation among different zones in Bangkok. Additionally, the study also analyzed factors affecting the level of household participation using Multiple Regression Analysis. Data were collected by means of hand-delivered questionnaires. A total of 400 respondents were selected using multi-stage random sampling by dividing Bangkok into three zones. The results showed that about two-thirds of the residents had got high level on knowledge and understanding on solid waste management. However, the results of ANOVA revealed that there was no significant difference between the level of household participation among residents who live in different zones. The level of participation in solid waste segregation and recycling of households in Bangkok was significantly influenced by promoting campaign and training programs continuously from local authorities and age of the residents. Finally, the discussion of the results of the study is presented and further study is also mentioned.

INTRODUCTION

Bangkok, the capital of Thailand is an enormous administrative area which has more than 1,500 square kilometers. At the 2010 census, Bangkok had overall total population of 8.28 million. Even though, in 2015 only about 6 million people were registered residents (<http://bangkok.go.th/info/>). Due to the large number of population, Bangkok like other big cities has faced a high level of environmental pollution and waste management problems. The capital produces about 9,900 tons of garbage daily or 1.53 kilograms per person but only 13% of waste is recycled per day. The remaining 8,700-plus tons are dumped in landfill (<http://thaipublica.org/2014/11/bangkok-big-garbage-problem/>). The main reasons of disposing of waste into landfill are that it is the simplest, cheapest and most cost-effective method (Barrett and Lawler 1995). Solid waste disposed in a landfill requires a complex process which also leads to hazardous emissions (Omar and Hani 2006). These have become a treat to human health and quality of life problems everywhere. A range of programs and policy instruments are required from the government and stakeholders to manage those waste appropriately in order to improve these problems. Creating an environmentally sustainable community requires an involvement of households in recycling solid waste (Kato et al. 2015). As such, it is necessary to increase the public awareness of waste generation and separation at source which will reduce the volume of waste to deposit. Not only waste reduction can help in reducing the expenditures and investment of government through lower collection, treatment and disposal but also protecting the environment.

Nowadays, environmental issues as well as the concern regarding the problems of waste have been increasing in every sector yet only little participation implements it. In order to promote recycling among households, it may require an understanding impact of factors affecting household participation in solid waste recycling. Consequently, the broad objective of this research is to investigate factors affecting household participation in solid waste recycling in Bangkok city, the capital of Thailand. Specifically, the research analyzes the level of community mobilization and knowledge on waste

management and examine the relationship among the level of community mobilization and knowledge of household participation and demographic characteristics towards the level of household participation in solid waste recycling.

The structure of the remainder of this paper is therefore organized as follows. Section 2 describes research methodology employed in this study. Results obtained from the survey are described in Section 3. Finally, conclusion, discussion as well as practical implications for policy makers and waste management planners are then discussed in the final section.

RESEARCH METHODOLOGY

Survey Design and Sampling Method

Data for this research were collected by means of hand-delivered questionnaires during November and December 2015. Population in this study referred to individuals residing in Bangkok, the capital of Thailand. The number of residents in Bangkok (N) was 5,696,409 people in 2015 (<https://th.wikipedia.org/>). Due to the population being enormous, a total of 400 sample sizes (n) were selected for this study (Yamane 1973). In order to define respondents who fill in the questionnaire, multi-stage sampling technique was used to select the respondents (Som 1996). At the first stage, stratified random sampling was employed by dividing Bangkok, which comprises of 50 districts, into three strata (h) i.e. inner area (21 districts), middle area (18 districts), and outer area (11 districts) as shown in Figure 1. Next, proportion allocation for each stratum (n_h) was assigned using $n_h = \frac{N_h n}{N}$ formula. Residents in each stratum were then selected using convenience sampling procedure. To obtain these samples, a questionnaire was distributed to each household by survey teams and finally 400 complete questionnaires were returned and used for further analysis.

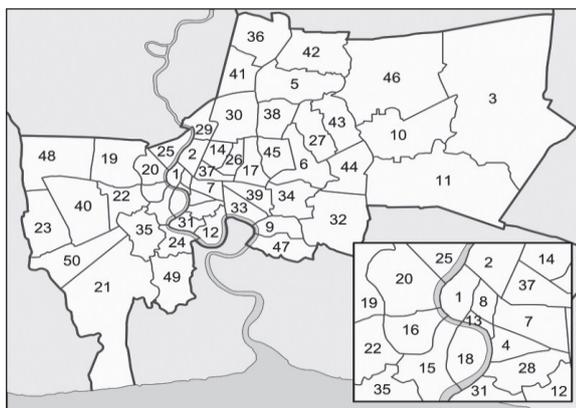


Figure 1: Bangkok Metropolitan Area

Source: https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok

Research Instruments

This research used a quantitative method and the instrument for gathering data was questionnaire, which was composed of five parts:

Part 1: Inquiries about general information, socio-economic characteristics of household and waste generation in the residence, as well as waste behavior.

Part 2: The respondents were asked whether they have knowledge and understanding about solid waste disposal and recycling.

Part 3: The questions regarding the level of community mobilization awareness in solid waste management such as adequate separated recycle bins advertisement and providing information concerns about solid waste management. Each question gauges according to a five-point Likert-type scale (Wolfer 2007). Level 1 means that factor was minimal supported in practices, whereas level 5 means that factor was supported at maximum in practices.

Part 4: The fourth part was the main part of the questionnaire including queries about the level of participation in waste recycling practices including 12 variables. Each variable gauges according to a five-point Likert-type scale (Wolfer 2007). Level 1 means that factor has minimal participation in waste recycling practices, whereas level 5 means that factor has maximum participating in waste recycling practices.

Part 5: Inquiries about opinions and ideas on how to promote households involving in solid waste recycling and source reduction, establish concerns and awareness in environmental issues and the existing barriers or obstacles for separating waste disposal in practices.

Reliability

A pilot study was carried out with 30 respondents. Cronbach's Alpha coefficient was used to test for a reliability of the instrument in Part 3 and 4 which were equal to 0.89 and 0.85, respectively. It is implied that the tool is sufficient and reliable for being used to collect data in primary source (Creswell 2002).

Analytical Techniques

Data analysis was performed using the Statistical Package for the Social Sciences (SPSS) version 19.0 for Windows. Descriptive statistics for each variable including mean, standard deviation and percentage were calculated to explain demographic characteristics, the level of household participation in solid waste recycling, the level of knowledge of waste management, types of waste generated at the residence, as well as methods of disposing solid waste in Bangkok. One-way Analysis of Variance (ANOVA) was employed to test whether there was statistically significance between the level of

household participation among three zones in Bangkok i.e. inner, middle and outer zones. Whilst Multiple Regression Analysis was employed to scrutinize factors affecting the level of household participation in solid waste recycling in the study. Before conducting regression analysis, correlations of all variables were investigated which were used to determine the bivariate relationships; for instance, the strength and direction between each predictor and dependent variable. The predictor variable which correlated to any of dependent variables were then further entered into Multiple Regression Analysis to explore the effect of variables on the level of household participation. Four principal assumptions of multiple regression analysis, i.e. linearity and additivity, statistical independence of the errors, homoscedasticity (constant variance) of the errors, and normality of the error distribution were tested before further using the models (Ryan 2009). The results of this research were accordingly displayed using statistical tables for interpretation in the following section.

The multiple regression model employed to investigate factors affecting household participation in solid waste segregation and recycling in this research is expressed as follows:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon_i \quad (1)$$

where:

- Y = The level of household participation in waste segregation and recycling practices
- X_1 = Gender (1 if women; 0 otherwise)
- X_2 = Age (1 if more than 35 years old; 0 otherwise)
- X_3 = Education (1 if undergraduate level or above; 0 otherwise)
- X_4 = Perceived suitable guidance on waste segregation from the municipality (1 if yes; 0 otherwise)
- X_5 = Knowledge of solid waste management
- X_6 = Promoting campaign and training programs continuously by local authorities
- X_7 = Creating network for protecting environment by communities
- X_8 = Providing enough different separate bins for encouraging segregation from local authorities
- ε_i = The random error term

RESULTS

Demographic Characteristics of the Respondents

According to the questionnaire, the respondents were more women (56.89%) than men (43.11%). The highest percent of respondent were 15-24 years and 25-34 years old (32.00%) equally, followed by 45-54 years old (16.75%). Half of the respondents had achieved undergraduate level (50.38%), followed by higher school certificate level (15.54%). 27.35% of the

respondents were working in private sectors and 22.03% are students/graduate students. 47.75% of the household size is 3-5 people, followed by more than 5 people (34.75%). The respondents indicated that almost half of major constitutes of wastes generate from households were general wastes e.g. plastic bags, foils, etc. (49.45%) and 34.34% are compostable wastes e.g. fruits and vegetables, leaves and so on.

The respondents were also asked how they disposed wastes at their residence. About 57% said that they did not segregate before disposal yet. The main reason or obstacle for not segregating solid waste at the residence was that the majority of respondents (36.48%) didn't have enough separation bins at source, followed the fact that even though they did not have any obstacles, they didn't intend to do it (34.07%). Whereas, the rest of the respondents (43%) claimed that they sort them at their residence before disposal. After segregation, most of them sold them for recycling at scrap dealers or garbage shops (55.61%). About 28% put them to the collection points after sorting in separate bags which was easy for collectors to further manage. When asked whether or not they had ever perceived awareness campaign on waste segregation and recycling in their communities from the municipality, 56.82% of the respondents indicated that they hadn't heard about it.

Knowledge on Solid Waste Management

Results from Table 1 showed that 68.74% of the residents had got high level on knowledge and understanding on solid waste management. The mean score of knowledge on solid waste management equals to 8.15 (S.D. = 1.24) which is quite high average score.

Table 1: Knowledge on Waste Management

Level of knowledge on waste management	n	Percent	(\bar{X})	S.D.
Low	4	1.03	3.75	0.50
Medium	117	30.23	6.75	0.54
High	266	68.74	8.84	0.72
Total	387	100.00	8.15	1.24

One-way Analysis of Variance (ANOVA)

Table 2 shows the results from using ANOVA to investigate the mean difference in the level of participation on solid waste segregation and recycling of household between residents who live in three zones in Bangkok namely inner, middle and outer areas. It was revealed that the level of participation on solid waste segregation and recycling of residents in three different zones were similar behaviour in nature.

Table 2: ANOVA Test

Source of Variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	308.100	2	154.050	1.532	0.217
Within Groups	39,929.490	397	100.578		
Total	40,237.590	399			

Factors Affecting Household Participation in Solid Waste Segregation and Recycling

Before conducting Multiple Regression Analysis, the assumptions were investigated. Table 3 displays the results of residual of normality test. Muticollinearity by examining tolerance and the Variance Inflation Factor (VIF) was also tested before further analysis and shown in Table 4. Furthermore, normal probability plot of the residuals in Figure. 2 shows that there is no data which stay far away from the slope line and the results from Table 5 also support the assumption. Hence, it can be implied that the regression model is appropriate for further study.

Table 3: The Result of Residual Normality Test

Residual Normality Test	Result
Standard Deviation	0.586
Kolmogorov Smirnov Z	0.038
Asymp.sig. (2-tailed)	0.183

Table 4: Muticollinearity and the Variance Inflation Factor (VIF) Values

Variables	p-value	Collinearity Statistics
Promoting campaign and training programs continuously by local authorities (X_6)	0.000**	1.001
Age (X_2)	0.050*	1.001

* significant at p-value < 0.05 level

** significant at p-value < 0.01 level

Table 5: Tests of Normality

	Kolmogorov-Smirnov*			Shapiro-Wilk		
	Statistic	df	p-value	Statistic	df	p-value
Standardized Residual	0.086	385	0.00**	0.983	385	0.00**

* Lilliefors significance correction

** significant at p-value < 0.01 level

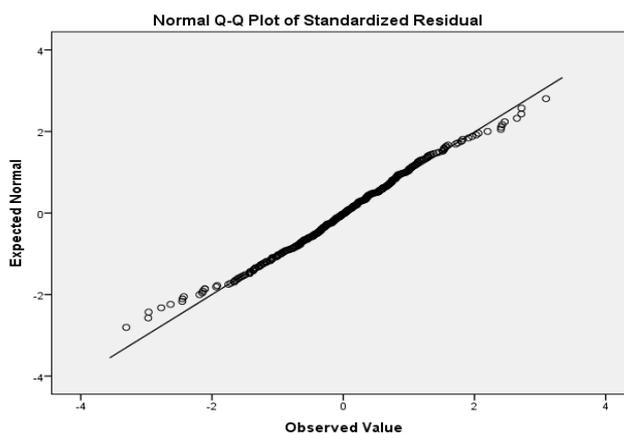


Figure 2: Graph of Normality Plot

Table 6 demonstrates the result of Multiple Regression Analysis explaining the variables which influence household participation in solid waste segregation and recycling. The result shows that promoting training programs continuously by local authorities and age have significant influence to the level of participation in solid waste segregation and recycling of households in Bangkok. The value of the coefficient of determination (R^2) indicates that 51.0% of the variations in the level of participation in solid waste segregation and recycling of household is explained in the regression model.

Table 6: Factors Affecting Household Pparticipation in Solid Waste Segregation and Recycling

Variables	Estimation	SE	t-statistics	p-value
X_6	0.598	0.03	20.029	0.000**
X_2	-0.121	0.06	-1.967	0.050*

* significant at p-value < 0.05 level

** significant at p-value < 0.01 level

$R^2 = 0.510$

F-value = 204.303

CONCLUSION

This research explored factors affecting the level of participation in solid waste segregation and recycling of households in Bangkok metropolis, as well as examining current Bangkok households waste management practices and their knowledge of waste management. Data were gathered from survey conducted using multi-staged sampling technique during November and December 2015 by means of hand-delivered questionnaires. It is interesting to note that the knowledge and understanding in solid waste management of the respondents of this study are at a high level. However, only 43% of them do segregate at their residence before disposal. The main reasons or obstacles for not segregating solid waste at home were that the majority of respondents (36.48%) didn't have

enough separation bins, followed by not intention to do so (34.07%). These findings imply that although the score of household's knowledge on recycling is high, the level of waste sorting is still low in practice. The result supports with the previous study of Latif et al. (2013) and Otitoju (2014). In addition, the findings is also in line with the research outcomes of Atthirawong (2015) which revealed that there was only 13.1% of the college students at KMITL university in Bangkok always separate their garbage before disposal even they had high level of opinion on solid waste management. The results from applying Multiple Regression Analysis indicated that there were only two factors which affected household participation of Bangkok residents' i.e. perceived promoting campaign or training program on solid waste management and age of the residents. As such, the local authorities should launch the intensive campaign, seminar orientation for waste education and awareness towards waste management continuously. The role of individual in waste management at source should be highlighted through regular media, campaign as well as consistent monitor (Ann et al. 2009). However, the findings indicate that age of the residents and the level of participation are in the opposite side. The evidence shows that the older people have lower participation in solid waste segregation and recycling. As such, those campaigns should target at both the older and the younger people continuously to encourage them in order to take part in segregation process in the future. It is necessary for each household to arrange a suitable space and separate bin for disposal in the residence to make an activity possible.

In order to make recycling to success, many parties are needed to be involved. Efforts are needed to engage by making collaboration and partnership with the relevant stakeholders. Policy formulation, developing efficient recycling initiatives and implementing an integrated waste management programs to the citizens will be needed to conduct and manage which should be started at the lowest level (Worku and Muchie 2012).

Enforcement people to practices waste recycling by laws, as well as determination different price levels of household disposal from waste generation may be compulsory to reduce the number of garbage collection and environmental issues. At the same time, it is crucial for municipal and communities to provide the necessary enabling facilities for the residents as well (Otitoju 2014). As Price (2001) mentioned, the role of local authority and public actions were prevailing to the success of sustainable waste policies. Thereby, it is also desirable to increase awareness of the environment problems to individuals as this can be changed individuals behaviors into everyday routine in waste recycling (Lober 2007). Incentive should also be contributed to the businesses, local authorities and municipal workers that in responsibility in collection and disposal of solid waste (Worku and Muchie 2012).

These issues should be implemented in the same direction.

Although this study has provided useful and update information on the level of participation of solid waste recycling of Bangkok citizen. There are still limitations of this study. A small of sample was undertaken compare to the number of population of the capital. This study also does not measure intention and attitude to segregate and recycle of the residents. Consequently, there are several issues to further investigation in these areas. For instance, the Theory of Planned Behaviour (TPB) framework (Ajzen 1991) should be employed for further investigation as the results of the study revealed that only knowledge on waste management as well as relevant factors may not be enough to enlighten behavior of the respondents in participation in solid waste management. The theory will enable to identify success key factors which are likely to encourage or discourage performance of behavior (Tonglet et al. 2004). The incorporation of additional variables would contribute to the explanation of the level of participation or behavior in recycling.

ACKNOWLEDGEMENT

The author would like to thank Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, the sponsor of this research funding. Appreciation also goes to all Bangkok residents who participated in this survey and all survey teams who assisted in administrating surveys.

REFERENCES

- Afroz, R., Hanki, K. and Tuddin, R. 2010. "The role of Socio-Economic Factors on Household Waste Generation: A Study in a Waste Management Program in Dhaka City. Banglades". *Research Journal of Applied Sciences*, Vol. 5, No.3,183-190.
- Ajzen, I. 1991. "The Theory of Planned Behaviour. Organizational" *Behavior and Human Decision Processes*, Vol. 50,179-211.
- Ann, A.P., Laurence, D., Charlemagne, F. and Armando, Jr., B. 2009. "Mobilizing Public Support for a Sustainable Solid Waste Management: The Case Study of Santo Tomas Municipality, Philippines". Institute for Global Environment Strategies (IGES) Policy Report.
- Atthirawong, W. 2015. "Opinion and Behaviour of Participation in Solid Waste Management of King Mongkut's Institute of Technology Ladkrabang's Students". *Business Review*, Vol.7, No.2, 41-58.
- Creswell, J.W. 2002. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. 2nd ed. SAGE Publications, Inc.
- Kato,T., Tran A.Q. and Hoang, H. 2015. "Factors affecting voluntary participation in food residue recycling: A case study in Da Nang, Vietnam Nam". *Sustainable Environment Research*, Vol.25, No.2, 93-101.
- Latif, S. A. et al. 2013. "Analyzing the Effect of Situational Factor on Recycling Behaviour in Determining the Quality of Life". *Journal of Asian Behavioral Studies*, Vol. 3, No.8, 37-46.

- Lober, D.J. 2007. "Municipal solid waste policy and public participation in household source reduction". *Waste Management Resources*, Vol.14,125-143.
- Rani, P. 2014. "Factors influencing consumer behavior". *International Journal of Current Research Academic Review*, Vol. 2, No.9, 52-61.
- Otitoju, T.A. 2014. "Individual Attitude toward Recycling of Municipal Solid Waste in Lagos Nigeria". *American Journal of Engineering Research*, Vol. 3, No.7, 78-88.
- Omar, A. and Hani, A. Q. 2006. "Municipal solid waste landfill citing for ecological science". *Waste Management*, Vol.26, 299-306.
- Price, J. L. 2001. "The landfill directive and the challenge ahead demands and pressures on the UK householder". *Resource Conservation and Recycling*, Vol. 32, No.3-4, 222-348.
- Ryan, T. P. 2009. *Modern Regression Methods*. 2nd ed. New Jersey, John Wiley & Sons, Inc.
- Som, R.K. 1996. *Practical Sampling Techniques*. 2nd ed. New York, Marcel Dekker Inc.
- Tonglet, M., Phillips, P.S. and Bates, M.P.2004. "Determining the drivers for householder pro-environmental behaviour: waste minimisation compared to recycling". *Resources, Conservation and Recycling*, Vol. 42, 27-48.
- Worku, Y. and Muchie, M. 2012. "An attempt at quantifying factors that affect efficiency in the management of solid waste produced by commercial businesses in the city of Tshwane, South Africa". *Journal of Environmental and Public Health*, 1-12. doi:10.1153/2012/165353.
- Wolfer, L. 2007. *Real Research: Conducting and Evaluating Research in the Social Sciences*. Boston, Pearson/Allyn and Bacon.
- Yamane, T. 1973. *Statistics: An Introductory Analysis*. 3rd ed. New York. Harper and Row Publications.
- <https://th.wikipedia.org/wiki/>. [available online] [access 31 March 2016].
- https://en.wikipedia.org/wiki/List_of_districts_of_Bangkok. [available online] [access 25 Jan 2016].
- <http://thaipublica.org/2014/11/bangkok-big-garbage-problem/>. [available online] [access 25 Jan 2016].
- <http://worldpopulationreview.com/world-cities/bangkok-population/> [available online] [access 25 Jan 2016].

AUTHOR BIOGRAPHIES



WALAILAK ATTHIRAWONG is Associate Professor of Operations Research at Faculty of Science, King Mongkut's Institute of Technology Ladkrabang (KMUTL) in Thailand. She had received Ph.D. in Manufacturing Engineering and Operations Management at the University of Nottingham in 2002. Her research interest are logistics and supply chain management, simulation, multi-criteria decision making and optimization. Her e-mail address is : walailaknoi@gmail.com

A plea for microsimulation

Marc Hannappel
Department of Sociology
University of Koblenz-Landau
56070, Koblenz, Germany
Email: marchannappel@uni-koblenz.de

KEYWORDS

Microsimulation; educational projections;

ABSTRACT

German demographic or educational projections are conventionally based on macrosimulations. Macrosimulation models use average values of the simulation parameters to compute the updating process. Additionally, the proportions of educational graduate rates stay constant in these models. In this paper I introduce a discrete event microsimulation model which is designed to project the graduate rates of the German population. With this simulation model it was demonstrated that the development of the graduation rates can become the subject of the simulation.

INTRODUCTION

Since the late 1950s, a similar development can be observed in all western industrial countries. This development can be described as an increase of the educational demand and, as a consequence, an increase of the educational level of the society (Hradil 2006; Allmendinger et al. 2010; Geißler 2008). The proportion of students with a lower educational attainment decreased between 1955 and 2000, whereas the proportion of students with an intermediate and high educational attainment increased (Allmendinger et al., 2010). Afterwards, the development can be described as a polarization of educational attainment. The proportion of students with an intermediate educational attainment stagnates whereas the development of the rest continues as the patterns before (Hannappel, 2015). In the political and scientific discourse, this development is usually called “the educational expansion”.

Germany occupies a middle rank in the PISA study (The Programme for International Student Assessment) (Deutsches PISA-Konsortium, 2001) which lead to a public and political controversy about the structure, the aims and the curriculum of the German education system. Hence, education policy becomes more and more important. Whereas the sociopolitical point of view focused on the development of the educational level of the population¹, the educational policy focused on the educational participation of the next generation, e.g. the transition rates within the education system and the graduation rates. Therefore, a main point for policymakers is to provide

¹A widespread assumption is that an increase of the educational level leads to an increase of prosperity and social innovation capability (Anger et al. 2006; Becker 1994; Bildungsberichterstattung 2006; Picht 1964).

an adequate educational infrastructure and a sufficient number of teachers (Klemm, 2012). However, educational reforms and teacher-training need time. Hence, education planning needs an anticipatory approach. To calculate the potential demand, policy makers need information about possible (or most likely) future developments of educational attainment. Educational projections of those developments are usually used as an information tool (Kultusministerkonferenz, 2002).

Problem: The German ‘Kultusministerkonferenz’ (KMK)² regularly calculates educational projections since 1963 (Kultusministerkonferenz, 2002). The projection of the graduates of the education system are not calculated completely by the KMK. The calculation is based on a population projection of the German Statistical Office (GSO). The GSO uses macrosimulation for the population projections (Statistisches Bundesamt, 2009). Macrosimulation models use mean values of certain parameters to project a system status (in the case of a population projection it is the age structure of a society) into the “future”. The parameters of the population projection are *fertility, migration* und *mortality*. The KMK uses these results to calculate the future graduates of the german education system (Kultusministerkonferenz, 2013). Additionally, the KMK uses current education transition rates and applies them to the calculated future population. At this point, however, there is a structural problem. The transition rates which are used for the calculation are percentage proportions of students who leave the school with a certain graduation. These proportions remain constant during the macrosimulation. Nevertheless, the previous development of the graduation rates never remain constant. From this point of view, the problem of this kind of projection is the structural separation of population projection and education projection.

Objectives: In this paper I present a simulation method which solves the problem of the separation. Instead of a macrosimulation model, I use a microsimulation model to project possible developments of educational attainment. Whereas an expansion of microsimulation models and techniques can be observed during the last decade (Li and O’Donoghue, 2013), a similar development is not recognizable for Germany.

Approach & Method: The projection of the educational attainment is based on an event-oriented dynamic microsimulation approach. Microsimulation models calculate single bio-

²The KMK is a conference of the education ministers of the different German federal states.

graphic events (e. g. *birth, partnership, education, death*, etc.) for every agent³. This method allows to implement different kinds of calculations, which determine time as an interval until an event will occur. All operations within the simulation are regarded as chronological sequences and every event occurs at a discrete time. The probability of single events can be calculated by conventional statistic methods. Furthermore, transition probabilities of a single event can be calculated depending on other events (independent variables), e.g. sex, age, educational status (Hannappel and Troitzsch, 2015). The calculations are based on macro-structural analyses, e.g. the calculation of the probability of women to give birth to a first child depending on age and educational status. This structure enables to analyze interaction effect between macro phenomena like the influence of demographic processes on educational attainment. Whereas the graduate rates remain constant in the macrosimulation model of the KMK, these rates are the subject of the microsimulation model.

Contributions: The construction of a microsimulation model needs time, manpower and money. As a consequence, the usage of microsimulation models is usually a privilege of larger (commercial) research institutes. The methodological discussions about different approaches and techniques takes place rather in the so called ‘method reports’ or ‘discussion papers’ than in scientific Journals (Axelrod 1997; Hannappel and Troitzsch 2015)⁴. This is unfortunate, because microsimulation models have a special analytical potential and can be used as an addition to conventional statistical methods. In particular, the current political situations make models necessary which are able to consider complex social phenomena in order to test possible future scenarios and to help policy makers make their decisions (Mannion et al., 2012).

Structure: Within microsimulation models, biographical events have to be modeled. It is necessary to have some knowledge about the factors the single events are influenced by. The first section discusses the main factors of the important events. Afterwards, I present the module structure and with the example of the event *first transition*, I will give an introduction to the operation mode of the simulation algorithm.

THEORY

As mentioned above, the microsimulation model, which will be discussed in the method section, cancels the separation between population projection and education projection. The microsimulation model considers educational processes and demographic processes. To construct a theory based model, assumptions from an educational theory as well as from demography theory are necessary. The main findings of the simulation are presented in the section “result” and discussed in the last section.

³Agents are simulated individuals or cases/actors from the starting dataset.

⁴Exeaptions are “The Journal of Artificial Societies and Social Simulation (JASS) <http://jasss.soc.surrey.ac.uk/JASSS.html> and the “International Journal of Microsimulation” <http://www.microsimulation.org/ijm/>

Educational sociology:

It is well known from numerous national and international studies that individual educational decisions are dependent on an individual’s social origin (Boudon 1974; Bourdieu 1982; Becker 2000; Becker 2010; Geißler 2005). The international comparison shows that the correlation between social origin and school performance is especially high for Germany (Deutsches PISA-Konsortium, 2001). The German education system is characterized by a horizontal stratification. After primary school the parents have to decide whether their children will go to a lower secondary, middle or higher secondary school. Basically, there are three main transitions: 1) the transition from primary school to secondary school, 2) transitions within secondary school (from one school type to another) and 3) the transition after secondary school to university or to vocational training or job.

1) The transition from primary school to secondary school is the most important transition for individual educational careers (Becker 2000; Becker and Lauterbach 2010; Ditton 2010; Henz and Maas 1995; Solga and Wagner 2010). This is mainly caused by two reasons: a) only a small number of students change the initially selected school type and b) a high correlation between educational graduation and occupational careers. 2) Transitions within secondary schools are rare; most transitions are downgradings (e.g. from high secondary schools to middle secondary schools) (Baumert et al. 2003; Solga and Wagner 2010). 3) Finally, a lot of studies show that the transition to university is dependent on the social origin as well (Middendorff et al. 2013; Müller et al. 2011).

Demography:

In the demographic research, two main factors have crystallized as important for fertility behavior: 1) region and 2) education.

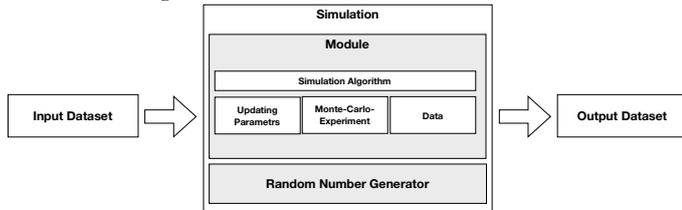
1) Even 25 years after the German reunification, there are stable differences in the fertility behavior between West(ern) and East(ern) Germany. Women from West-Germany give birth to their first child later, have a higher rate of childlessness and generally have a lower number of children (Kreyenfeld and Konietzka, 2004).

2) Furthermore, it is well-known that the probability to give birth to a child varies between women with different educational status. The higher the educational level of women, the lower the probability to give birth to a child (Blossfeld et al. 1991; Brüderl and Klein 2003; Herlyn et al. 2002; Klein and Lauterbach 1994; Wirth 2007). Other factors like class, employment and religious denomination are less important (Höpflinger, 2012)

These findings were considered by the construction of the microsimulation model⁵.

⁵The points listed above are rather empirical than theoretical findings. However, this paper is not the place for a detailed reception of the theoretical explanations. Boudon (1974) and Bourdieu (1982) give a deep insight into the relation between the social phenomena which are discussed above.

Fig. 1. Elements of microsimulation models



METHOD

Microsimulation:

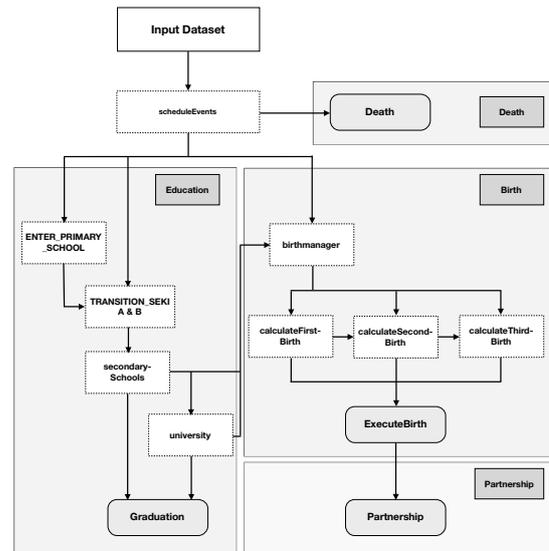
Figure 1 shows the main elements of a microsimulation. At first, a dataset is necessary. During the simulation, the dataset will be updated to a certain time point (t_T) depending on decision rules which are implemented in the model (Gilbert and Troitzsch 2005; Spielauer 2009). Dynamic microsimulation models can either be modelled with a period approach or with an event-orientated approach. Whereas the former ones model time as a process between determined time intervals (mostly one year), event oriented simulations model time as a random variable which is simulated as the time between two events. In period-oriented models, the simulation runs from one time step to another. In event-oriented models, the simulation runs from event to event. The events like *birth*, *school enrolment*, *marriage*, etc. are organized in single simulation modules. The main item of a simulation module is the simulation algorithm. Simulation algorithms are formalized program codes which structure the order of the modules. In combination with Monte-Carlo-Experiments (Galler 1997; van Immhoff and Post 1998), the simulation algorithms decide whether an event occurs to an agent or not. Finally, the result of a simulation model is a fictitious dataset which includes the simulated population. The advantage of microsimulation is that the outcome of these models has the same structure as the input dataset. Consequently, the dataset can be analyzed with the conventional statistical methods and software.

The model:

Figure 2 gives an overview of the model structure. I used the Scientific Use File (SUF) 2008 as the input dataset. The SUF is a 70% subsample of the microcensus which includes 477 239 individuals. The microcensus is a 1% annual census of the German population which includes a lot of questions about socio-economic issues. The simulation starts with a *scheduleEvent*. All actors from the starting population and all agents⁶ have to pass this event. Within this event, a simulation algorithm checks the characteristics of the agents and decides what will happen next to the agent.

In contrast to conventional statistical methods, microsimulation models are more complex. The description of the single

Fig. 2. Overview: Modules



modules, the simulation algorithms, the calculation of the transition rates and the description of the different datasets, which are the basis of the calculations, is too complex for this paper. Hence, I will describe the single modules in a simplified way and will then describe how the simulation works by the example of a single event⁷.

Death: The module *Death* is the first module the individuals have to pass. The simulation algorithm calculates the exact time of death for every individual. The time of death is a result of a comparison between a survival function value which is taken from the life tables of the German Statistical Office and a random number, which is drawn for every individual, i. e. the calculation of the date of death is the result of a Monte-Carlo-Experiment.

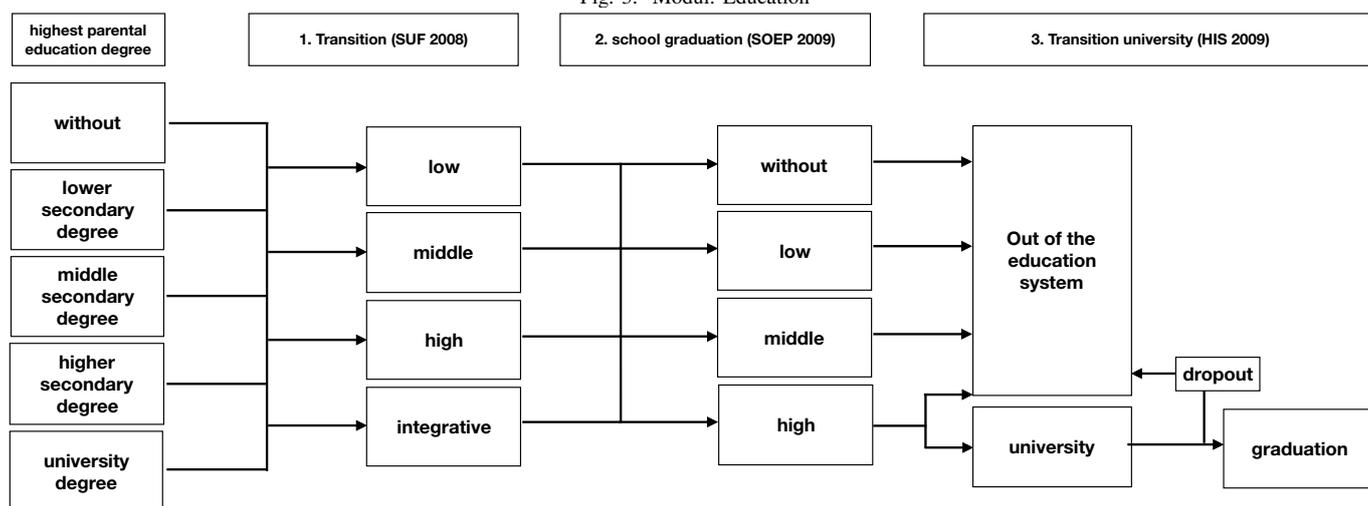
Education: Within the module *Education*, the simulation algorithm decides which kind of educational attainment will be assigned to an agent considering his or her “parents”. This module takes into account the different social selectivity mechanisms within the German education system (see the section ‘Education model’).

Birth: The module *Birth* consists of three different sub-modules (*calculateFirstbirth*, *calculateSecondbirth* and *calculatethirdbirth*). The submodule *calculateFirstbirth* calculates the day when a certain women will give birth to her first child depending on her educational status, age and whether she lives in West or East Germany. The submodule *calculateSecondbirth* and *calculateFirstbirth* calculates the day of birth of the second and the third child as well. However, whereas the first birth is calculated as the age of the women by her first birth, the second and third birth is calculated as the distance between

⁶The input dataset includes information from real persons. These cases are called ‘actors’. During the simulation the simulation model creates new cases (e. g. if a “child” is born). Cases which are created by the simulation are called ‘agents’

⁷For a detailed description of the complete simulation see (Hannappel, 2015)

Fig. 3. Modul: Education



first and second and second and third birth⁸.

Partnership: Finally, when a woman gives birth to a child, a partner is matched by education and age. As a consequence, fertility is modeled independently of partnership. This approach is oriented towards an approach that Martin Spielauer (2003) used in a model for the Austrian society.

The module education:

Figure 3 shows the empirical model of educational decisions. The education module considers three transitions. The first transition is the transition from primary to secondary school. The second transition calculates the school graduations of the agents and the third transition is the transition to university. This structure is based on the main empirical findings from the educational sociological research.

The first transition is calculated depending on the social origin of the students. I calculated the probability of a transition decision depending on the highest parental educational degree. For the calculation I used the SUF 2008. This dataset is used for the simulation as well. The school graduation is calculated depending on the school type of the secondary school in Germany. Principally, it is possible to catch up a school graduation. Therefore, I calculated the probability to get a certain graduation depending on the school form the students visit during secondary school. For those calculations, a panel dataset is necessary. Hence, I used the Socioeconomic panel (SOEP) from the German Institute of Economics (DIW) (Siegel et al., 2010). As mentioned above, the transition to a university is highly correlated with the social origin of the students. For this calculation, I used another panel dataset, the “Studienberechtigtenpanel” (Spangenberg et al., 2011) from the Hochschulinformationssystem (HIS). With this dataset, it

is also possible to calculate the probability of the transition to university depending on the social origin of the students.

Example: the first transition

Table I lists the probabilities for the first transition depending on the social origin of the students. The probability for children from parents without an education degree to visit an integrative school⁹ amounts to 10,4%. It is also shown that the

TABLE I
TRANSITION RATES OF THE FIRST TRANSITION DEPENDING ON THE HIGHEST PARENTAL EDUCATION DEGREE (IN PERCENT)

school type	highest parental education degree				
	Without	low	middle	high	university
integrative	16,1	10,4	10,2	10,3	8,8
high	6,3	11,3	27,6	45,2	67,3
middle	22,3	34,6	43,4	31,5	18,6
low	55,3	43,6	18,8	43,6	5,3
N	415	3248	4766	1722	2875

Source: FDZ der Statistischen Ämter des Bundes und der Länder, Scientific Use File 2008, own calculations, weighted with EF952, ²unweighted

transition rates vary between the parental education categories. The higher the parental education degree, the higher the probability to visit a high school type. Using the example of students from parents without an education degree, it will be shown how the simulation algorithm of the first transition works.

The simulation algorithm operates with integers between 0 and 100 000. Therefore, the percentage values have to be converted into integer values. Table II lists the percentage values (for students from parents without an education degree), the cumulative frequencies, the values for 1 - the cumulative

⁸The calculation method is described in a previous paper of the ECMS conference (Hannappel et al., 2012), see also (Hannappel and Troitzsch, 2015) or (Hannappel, 2015).

⁹Integrative schools in Germany are schools which integrate all school types, i. e. students from this school type can reach all kinds of school leaving qualifications.

frequencies and the simulation values. The simulation values are the values vor 1 - the cumulative frequencies multiplied by 10 000. These are the values which are used in the simulation.

TABLE II
CALCULATION OF THE SIMULATION VALUES

school type	Percent	Cum. Frequencies	1-Cum. Frequencies	Simulation Values
integrative	10,4	10,4	0,9	100 000
high	11,3	21,8	0,78	89 569
middle	34,6	56,4	0,44	78 247
low	43,6	100	0	43 628
N^2	3248			

Source: FDZ der Statistischen Ämter des Bundes und der Länder, Scientific Use File 2008, own calculations, weighted with EF952, ²unweighted

First of all, the random number generator draws a random number between 0 and 100 000. The simulation algorithm compares this random number with the number of the second row from the column “Simulation”. If the random number is larger than 89 569, the agent will sign as a student on an integrative school form. Otherwise the random number will be compared with the value of the third row (category “middle”). If the random number is larger than 78 247, the agent will sign as a student on a high school type. This process will be repeated until a suitable category is found. This approach is called “Monte-Carlo-Simulation” and is based on the probabilistic assumption that a large number of Monte-Carlo-Experiments leads to an approximation of the empirical values to the simulated cases.

$$\lim_{n \rightarrow \infty} P(|p(A) - P(A)| \leq \varepsilon) = 1 \quad (1)$$

(Kühnel and Krebs, 2001, S. 132 f.)

The transitions of agents from the category “parents without an education degree” during the simulation should be similar to the implemented values from table II.

Verification

The probabilistic design of microsimulation models leads to the problem that the results of microsimulations vary from simulation run to simulation run. Additionally, prospective simulations can not be validated¹⁰. One possibility to test the correctness of the simulation model is to verify¹¹ the model. Hence, to verify the model, the divergence of the output values from the input values have to be analyzed. Figure III shows the results of the chi-square test.

Besides only a few exceptions, the chi-square test shows that the model works in a desirable way. The main modules

¹⁰The only possibility to validate prospective simulations is to wait until the real time has reached the end of the simulation time to compare the simulation results with the real development. It is clear that this is no constructive approach.

¹¹Verification in the context of simulation models means the “process of checking that a program does what it was planned to do” (Gilbert and Troitzsch, 2005).

TABLE III
VERIFICATION: CHI-SQUARE-TEST & R^2

		χ^2	df	Sig	R2		χ^2	df	Sig	R2	
Bildung	OA	4,6	4	0,33	0,998	Geburten (West)	OA	3,8	3	0,28	0,998
	HS	4,4	4	0,35	0,999		HS	1,2	3	0,75	0,999
	RS	7,1	4	0,13	0,999		RS	4,6	3	0,20	0,999
	Abi	0,2	4	0,71	0,999		Abi	4,3	3	0,23	0,998
	Uni	0,4	4	0,63	0,999		Uni	6,9	3	0,07	0,996
Partnerschaft	HS	195,9	4	0,00	0,997	Tod	Männer	89,25	100	0,77	0,999
	RS	620,2	4	0,00	0,930		Frauen	103,6	100	0,28	0,999
	Abi	280,5	4	0,00	0,952						
	Uni	170,8	4	0,00	0,999						

(*Birth and Education*) show (nearly) no significant deviations. The goodness of fit (R^2) is almost above 90 % and mostly above 99 %. Only the module *partnership* shows significant deviations¹²

RESULTS

Figure 4 shows the development of the real educational attainment until the year 2012 and the results of the simulation for 2013 to 2050. The simulation results are based on the analysis of the agents from the age group from 26 – 35 years. Because of the low number of cases the analyses of the ALLBUS dataset are based on the 30 – 40-year-old population. The age when the agents get a certain graduation depends on the simulated educational career. Agents with a low school career get their graduation at the age of 15, with a middle graduation at the age of 16, with a high graduation at the age of 19 and with a university graduation at the age of 25. Therefore, the analysis of the simulation results can only be calculated for the agents from the age of ≥ 26 .

The simulation results show a plausible development. The proportion of agents with at least a high graduation increases from 42,5% to 46,6%. Hence, the simulation continues the educational expansion. Interestingly, the proportion of agents without a school graduation remain constant over the simulation time. This is remarkable because the population without a school graduation is characterized by a very high birth rate. The results show that different fertility rates are overcompensated by educational mobility.

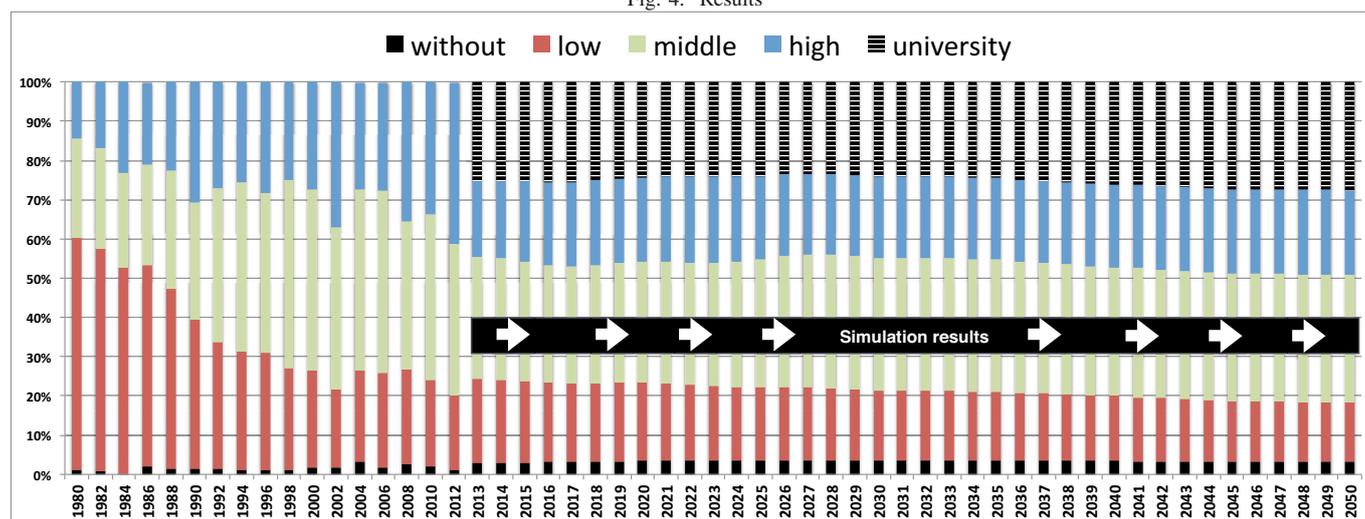
The development until the year 2025 may seem very plausible at the first sight. However, the reduction of the percentage of agents with at least a high graduation needs an explanation. The problem is caused by a problematic heuristic for the calculation of the transition for people in vocational training. These students were handled like students on an integrative school. This leads to an overestimation of the agents with a high graduation at the beginning of the simulation. This needs to be corrected in further simulation models. Therefore, new analyses with other datasets are necessary.

Despite the slight bias, the simulation results show:

- A continuation of the educational expansion in the upcoming years.

¹²Further analysis with percentage deviations, which cannot be described in this paper, shows that the differences between implemented values and the simulation results are very small (Hannappel, 2015).

Fig. 4. Results



Source: 1980 – 2012: ALLBUS 1980 – 2012 (adjusted version of the variable v668), 2013 – 2050: simulation results

- However, the increase of this expansion will lose momentum.
- The expansion is largely constituted by educational mobility.

CONCLUSIONS

Conventionally, macrosimulations are used to project possible developments of the educational attainment (Kultusministerkonferenz 2002; Kultusministerkonferenz 2013). Macrosimulation models use average values of the simulation parameters to compute the updating process. This prevents the possibility of interaction effects. Additionally, the proportions of educational graduate rates stay constant in these models.

With this simulation model it could be demonstrated that the development of the graduation rates can become the subject of the simulation. Although the transition rates remain constant in this model as well (indeed on a more detailed level), the graduate rates did not. The interaction between demographic educational parameters (education model & transition rates) leads to a variation of the graduation rates during the simulation. This is the main advantage of microsimulation over macrosimulation.

Especially in times of great changes, such models are helpful in order to get a better understanding for complex developments and contexts. It would be regrettable to renounce these models.

REFERENCES

Allmendinger, J., Ebner, C., and Nikolai, R. (2010). Soziologische Bildungsforschung. In Tippelt, R. and Schmidt, B., editors, *Handbuch Bildungsforschung*, pages 47–70. VS Verlag, Wiesbaden, 3 edition.

Anger, C., Plünn, A., Seyda, S., and Werner, D. (2006). *Bildungsarmut und Humankapitalschwäche in Deutschland*. Institut der deutschen Wirtschaft, Köln.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In Conte, R., Hegselmann, R., and Terna, P., editors,

Simulating social Phenomena, Lecture Notes in Economics and Mathematical Systems, pages 21–40. Springer, Berlin.

Baumert, J., Trautwein, U., and Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. In PISA-Konsortium Deutschland, editor, *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*, pages 261–331. Leske + Budrich, Opladen.

Becker, G. S. (1994). *Human Capital: Theoretical and empirical Analysis with special Reference to Education*. University of Chicago Press, Chicago and London, 3 edition.

Becker, R. (2000). Klassenlage und Bildungsentscheidungen. Eine empirische Anwendung der Wert-Erwartungstheorie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 52(3):450–474.

Becker, R. (2010). Soziale Ungleichheit von Bildungschancen und Chancengerechtigkeit – eine Reanalyse mit bildungspolitischen Implikationen. In Becker, R. and Lauterbach, W., editors, *Bildung als Privileg. Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit*, pages 161–189. VS Verlag, Wiesbaden, 4 edition.

Becker, R. and Lauterbach, W., editors (2010). *Bildung als Privileg: Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit*. VS Verlag, Wiesbaden.

Bildungsberichterstattung, K. (2006). *Bildung in Deutschland. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration*. W. Bertelsmann Verlag, Bielefeld.

Blossfeld, H.-P., Huinink, J., and Rohwer, G. (1991). Wirkt sich das steigende Bildungsniveau der Frauen tatsächlich auf den Prozess der Familienbildung aus? Eine Antwort auf die Kritik von Josef Brüderl und Thomas Klein. *Zeitschrift für Bevölkerungswissenschaft*, 17(3):337–351.

Boudon, R. (1974). *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society*. John Wiley & Sons Inc, New York.

Bourdieu, P. (1982). *Die feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft*. Suhrkamp, Frankfurt a. M.

Brüderl, J. and Klein, T. (2003). Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland, 1960–2000: Eine empirische Untersuchung mit dem Familiensurvey 2000. In Bien, W. and Marbach, J. H., editors, *Partnerschaft und Familiengründung*, volume 11 of *DJI: Familien-Survey*, pages 189–218. Leske + Budrich, Opladen.

Deutsches PISA-Konsortium (2001). *Pisa 2000. Basiskompeten-*

- zen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich, Opladen.
- Ditton, H. (2010). Der Beitrag von Schule und Lehrern zur Reproduktion von Bildungsungleichheit. In Becker, R. and Lauterbach, W., editors, *Bildung als Privileg*, pages 243–271. VS Verlag, Wiesbaden.
- Galler, H. P. (1997). Discrete-time and continuous-time approaches to dynamic microsimulation reconsidered. Technical Report 13, National Centre for Social and Economic Modelling, Canberra. <http://www.natsem.canberra.edu.au/storage/tp13.pdf> (13.03.2015).
- Geißler, R. (2005). Die Metamorphose der Arbeitertochter zum Migrantensohn. Zum Wandel der Chancenstrukturen im Bildungssystem nach Schicht, Geschlecht, Ethnie und deren Verknüpfungen. In Berger, P. A. and Kahlert, H., editors, *Institutionalisierte Ungleichheiten*, Bildungssoziologische Beiträge, pages 71–100. Juventa, Weinheim.
- Geißler, R. (2008). *Die Sozialstruktur Deutschlands: Zur gesellschaftlichen Entwicklung mit einer Bilanz zur Vereinigung*. VS Verlag, Wiesbaden, 5 edition.
- Gilbert, G. N. and Troitzsch, K. G. (2005). *Simulation for the Social Scientist*. Open University Press, Maidenhead, Berkshire, New York, 2 edition.
- Hannappel, M. (2015). *(Kein) Ende der Bildungsexpansion in Sicht?! Ein Mikrosimulationsmodell zur Analyse von Wechselwirkungen zwischen demographischen Entwicklungen und Bildungsbeteiligung*. Metropolis, Marburg.
- Hannappel, M. and Troitzsch, K. G. (2015). Mikrosimulationsmodelle. In Braun, N. and Saam, N. J., editors, *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, pages 455–489. Springer VS, Wiesbaden.
- Hannappel, M., Troitzsch, K. G., and Bauschke, S. (2012). Demographic and educational projections. building an event-oriented microsimulation model with comics ii. In Troitzsch, K. G., Möhring, M., and Lotzmann, U., editors, *ECMS 2012*. Pirrot, Dudweiler.
- Henz, U. and Maas, I. (1995). Chancengleichheit durch die Bildungsexpansion. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 47(4):605–633.
- Herlyn, I., Krüger, D., and Heinzlmann, C. (2002). Später erste Mutterschaft – erste empirische Befunde. In Schneider, N. F. and Matthias-Bleck, H., editors, *Elternschaft heute*, Zeitschrift für Familienforschung Sonderheft 2. Leske + Budrich, Opladen.
- Höpfinger, F. (2012). *Bevölkerungssoziologie: Eine Einführung in demographische Prozesse und bevölkerungssoziologische Ansätze*. Beltz Juventa, Weinheim, 2 edition.
- Hradil, S. (2006). *Die Sozialstruktur Deutschlands im internationalen Vergleich*. VS Verlag, 2 edition.
- Klein, T. and Lauterbach, W. (1994). Bildungseinflüsse auf die Heirat, die Geburt des ersten Kindes und die Erwerbsunterbrechung von Frauen: Eine empirische Analyse familienökonomischer Erklärungsmuster. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 46(2):278–298.
- Klemm, K. (2012). Zur Entwicklung des Lehrkräftebedarfs in Rheinland-Pfalz. *Bildungsforschung Bildungsplanung*.
- Kreyenfeld, M. and Konietzka, D. (2004). Angleichung oder Verfestigung von Differenzen? Geburtenentwicklung und Familienformen in Ost- und Westdeutschland. *MPIDR Working Paper*, (2004-025).
- Kühnel, S.-M. and Krebs, D. (2001). *Statistik für die Sozialwissenschaften: Grundlagen, Methoden, Anwendungen*. Rowohlt, Reinbek.
- Kultusministerkonferenz (2002). *Vorausberechnung der Schüler- und Absolventenzahlen 2000 bis 2020*. Sekretariat der Kultusministerkonferenz, Bonn.
- Kultusministerkonferenz (2013). *Vorausberechnung der Schüler- und Absolventenzahlen 2012 - 2025*. Sekretariat der Kultusministerkonferenz, Berlin.
- Li, J. and O'Donoghue, C. (2013). A survey of dynamic microsimulation models. uses, model structure and methodology. *International Journal of Microsimulation*, 6(2):3–55.
- Mannion, O., Lay-Yee, R., Wraposn, W., Davis, P., and Pearson, J. (2012). Jamsim: a microsimulation modelling policy tool. *Journal of Artificial Societies and Social Simulation*, 15(1):1–14.
- Middendorff, E., Apolinarski, B., Poskowsky, J., Kandulla, M., and Netz, N. (2013). *Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2012. 20. Sozialerhebung des Deutschen Studentenwerks, durchgeführt durch das HIS-Institut für Hochschulforschung*. Bundesministerium für Bildung und Forschung, Berlin.
- Müller, W., Pollak, R., Reimer, D., and Schindler, S. (2011). Hochschulbildung und soziale Ungleichheit. In Becker, R., editor, *Lehrbuch der Bildungssoziologie*, pages 289–327. VS Verlag, Wiesbaden, 2 edition.
- Picht, G. (1964). *Die deutsche Bildungskatastrophe. Analyse und Dokumentation*. Walter-Verlag, Olten and Freiburg.
- Siegel, N. A., Huber, S., Gensicke, A., Geue, D., Stimmel, S., and Stutz, F. (2010). Soep 2009. Methodenbericht zum Befragungsjahr 2009 (welle 26) des Sozio-ökonomischen Panels. Berlin: Deutsches Institut für Wirtschaftsforschung.
- Solga, H. and Wagner, S. (2010). Die Zurückgelassenen – die soziale Verarmung der Lernumwelt von Hauptschülerinnen und Hauptschülern. In Becker, R. and Lauterbach, W., editors, *Bildung als Privileg. Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit*, pages 191–219. VS Verlag, Wiesbaden, 4 edition.
- Spangenberg, H., Beuße, M., and Heine, C. (2011). Nachschulische Werdegänge des Studienberechtigtenjahrgangs 2006. Dritte Befragung der studienberechtigten Schulabgänger/innen 2006. 3 1/2 Jahre nach Schulabschluss im Zeitvergleich. *HIS: Forum Hochschule*, (18).
- Spielauer, M. (2003). *Family and education: intergenerational educational transmission within families and the influence of education on partner choice and fertility ; analysis and microsimulation projection for Austria*, volume 11 of *Schriftenreihe / Österreichisches Institut für Familienforschung*. ÖIF - Österreichisches Institut für Familienforschung, Wien.
- Spielauer, M. (2009). Microsimulation approaches. Ottawa: Statistics Canada – Modeling Division. <http://www.statcan.gc.ca/eng/microsimulation/modgen/new/chap2/chap2> (13.03.2015).
- Statistisches Bundesamt (2009). *Bevölkerung Deutschlands bis 2060: 12. koordinierte Bevölkerungsvorausberechnung*. Statistisches Bundesamt, Wiesbaden.
- van Immhoff, E. and Post, W. (1998). Microsimulation for population projection. *Population: An English Selection*, 10(1):97–138.
- Wirth, H. (2007). Kinderlosigkeit von hochqualifizierten Frauen und Männern im Paarkontext- eine Folge von Bildungshomogamie? In Konietzka, D. and Kreyenfeld, M., editors, *Ein Leben ohne Kinder*, Springer-11776 /Dig. Serial]. VS Verlag für Sozialwissenschaften — GWV Fachverlage GmbH Wiesbaden, Wiesbaden.

AUTHOR BIOGRAPHIES

Marc Hannappel was born 1980 in Bad Schwalbach, Germany. He studied educational research at the University of Koblenz-Landau and obtained his degree 2006. From 4/2008 to 9/2010 he was a scholarship holder of the Hans-Böckler-Stiftung. 2015 he finished his PhD thesis. Since 10/2010 he has been working at the Institute of Sociology at the University of Koblenz-Landau. His email is marchannappel@uni-koblenz.de.

SMALL SHIPMENT DELIVERY'S QUALITY IMPROVEMENT IN CITIES WITH UNSTABLE TRAFFIC

Pavels PATLINS

Faculty of Engineering Economics and Management
Riga Technical University,
1st Kalku Street, Riga, Latvia
pavels.patlins@rtu.lv

Remigijs Pocs

Faculty of Engineering Economics and Management
Riga Technical University,
1st Kalku Street, Riga, Latvia
Remigijs.Pocs@rtu.lv

KEYWORDS

Small shipments, delivery problem in cities (DPC), minimal-growth method, circular route,

ABSTRACT

The paper deals with small shipment delivery planning problems for cities with hard and unstable or intensive traffic. The problem is very significant today especially for big cities. When planning road transportation and routes within a big city it is expedient to work out the optimal system and achieve the best result. The most important criteria is a ratio customers' demand satisfaction – it is not enough to plan minimal transportation costs or vehicle's run only.

Total delivery time as well as a rate of accuracy often is a corner-stone factor, which influence the quality of delivery. It is necessary to choose the best order how to serve customers into each circular route.

The authors divide small shipment delivery problem into two groups – “problematic” and “non-problematic” routes planning (depending on the traffic intensity in the particular city). The authors of the paper recommend using of the minimal growth method (MGM) to improve vehicle route planning problems especially in cities with intensive traffic.

INTRODUCTION

Local deliveries planning specificity connected with the various restrictions as well as with the fact that small-quantity loading and transportation management is different from full-cargo transportation.

Actually, small shipment delivery problem is extremely important step into the total logistics chain (figure 1). Intercity (also international transportation) usually managed as simply-scheme FTL (full truck load) transportation, whereas local transportation often may be planned using circular route scheme.

Circular routes connect more than 2 points within one route, namely, the forwarder should take cargo from the consignor and deliver small quantity of cargo to a great number of recipients (multi-drop route).

This is daily problem for forwarders, trading

companies and manufacturers in big cities.

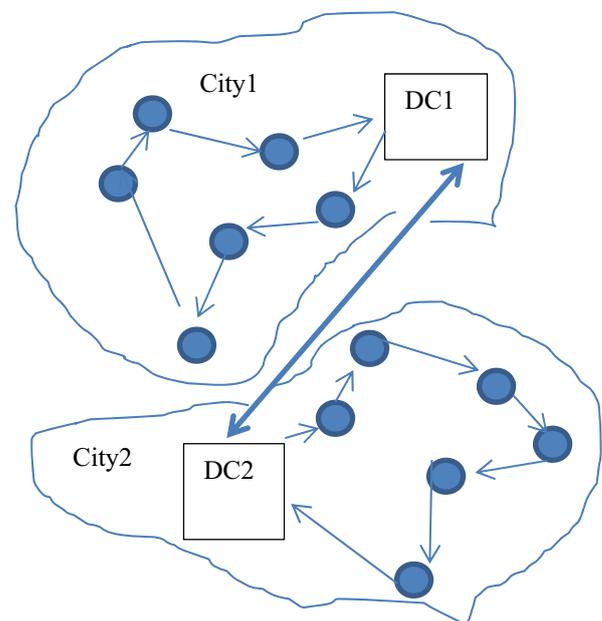


Figure 1. Intercity and city transportation's scheme.
DC1- Distribution centre of the first city (e.g. – wholesaler's warehouse);
DC2- Distribution centre of the second city (e.g. – wholesaler's warehouse);

↔ Intercity transportation;
← Transportation in cities.

The total delivery time on the j multi-drop route (from j -consignor to the customers assigned to him) is determined by using the formula:

$$T_{Tj} = t_{Lj} + \sum_{i=1}^k (t_{LDij} + t_{UNij}) + t_{Rj}, \quad (1)$$

where t_{Lj} - j -consignor loading time, h; t_{LDij} - the time while driving car on the i -length of j -multi-drop route (from $i-1$ to i -route point, where zero point is the depot or loading point), h; t_{UNij} - unloading time at i -consumer on j -multi-drop route, h; k - the number of unloading points on j -multi-drop route; t_{Rj} - return journey time to j -multi-drop route, h.

It should be noted that the delivery time depends not only on the forwarders' planning, but also on the management system of suppliers and consumers, particularly on their schedule (the number of breaks, length of dinnertime, etc.). The logistics approach of time modelling for transportation services necessitates the coordination between the motor transport and working schedule of the suppliers and cargo consumers, talking about the JIT(just-in-time) fulfilment of motor transport contractual obligations to the customers and suppliers. So, the corner-stone of planning is to determine the delivery time of the daily load just in time. So, the vehicle initial time may be determined, using following formula:

$$T_{in} = T_{JIT} - \sum_j T_{oj} - T_{i.r.}^1, \quad (2)$$

where T_{JIT} – the time of delivery of the “contractual” volume of goods JIT, h;

$T_0 = \sum_j T_{oj}$ – consumer goods delivery total time, h;

$T_{i.r.}^1$ – idle running time (from the motor carrier to the first point of loading),h.

All components of the formula 2 are random values. While determining the total delivery time on j-multi-drop route and realizing statistic modelling, it is necessary to take into account the work of supplier and a consumer, in particular, start and end time of the technological breaks of clients. So, the formula 1 should be transformed as follows:

$$T_{oj} = t_{Lj} + \sum_{i=1}^k (t_{LDij} + t_{UNij}) + t_{Rj} + \eta_j + \sum_{i=1}^k \psi_{ij}, \quad (3)$$

where η_j – the random component, taking into account j-supplier's technological breaks, h; $\sum_{i=1}^k \psi_{ij}$ – random component, taking into account j-supplier's technological breaks assigned to j-consumer, h.

Thus, in a real life it is a difficult and complex task to plan accurate deliveries in a big city with intensive traffic.

There are many special methods and algorithms to optimize circular route planning and reduce delivery costs as well as vehicle's run and transportation costs. Various authors investigated it. Often it is impossible to provide the needed result, because use of these methods is connected with the following problems:

- it is impossible to provide accurate result;
- it is impossible no satisfy customers' individual needs;
- it is impossible to take into account traffic intensity changes depending on days of the week and hours of a day.

The authors suppose that various computer programs often provide non-optimal solution to transport problem due to different restrictions; on the one hand, it is possible to use only these programs to solve theoretical problems.

On the other hand the real situation is changing daily,

because demand is not stable, it is possible to use heuristic methods. But in a real life many computer programmes are used in combination with heuristic method to achieve the optimal result. Because of lack of information, criteria of optimization used in practical conditions often are vehicle's run and transportation costs, not delivery time factor.

REVIEW OF LITERATURE

Transport problem is a well-known network optimization problem, first created by F.Hitchkok (1941). The goal was to find the optimal costs of distribution plan for one product delivery, multiplying it with quantity of product to find each channel and source capacity for each recipient.

When transportation costs of the given route are non-linear dependent on the quantity of production for transportation, this problem becomes a non-linear transportation problem. To find the optimal solution for this problem (NTP), it is necessary to make many investigations in logistic management. Many heuristic methods as well as mathematical program methods are created to solve NTP problems.

Many authors and researchers have worked out different methods and algorithms to solve NTP. As regards approximate heuristic optimization methods, genetic algorithms (GA) by Holland (1975), tabu search (TS) by Glover (1977), particle swarm optimization (PSO) by Kennedy and Eberhart (1995). Many specialists solve NTP, using also linear programming models. For instance, Cao (1992), Dangalchev (1996), Bell et al. (1999), Kuno and Utsunomiya (2000), Dangalchev (2000) and Nagai and Kuno (2005). However, research effort has been also devoted to nonlinear programming (NLP) techniques for the optimum solution of the NTP. For instance, Michalewicz et al. (1991) have applied the reduced gradient (RG) method to obtain the optimal solution of the NTP.

NEW CONCEPTS OF SMALL SHIPMENT HIGH-QUALITY DELIVERY'S PLANNING IN CITIES WITH UNSTABLE TRAFFIC

The authors divide transport network optimization models using the following classification (figure 1):

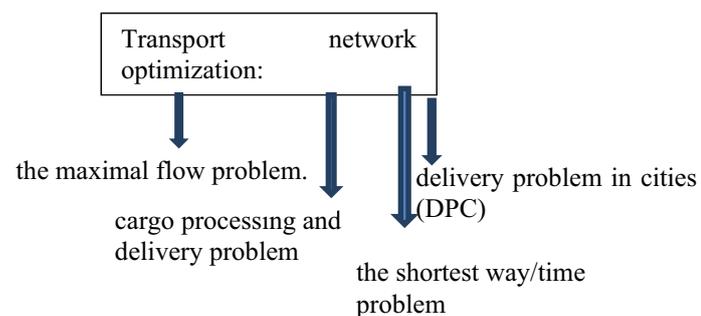


Figure 2. Transport network optimization models.

So, the main transport network optimization problems are:

- a) cargo processing and delivery problem;
- b) the shortest way/time problem;
- c) the maximal flow problem.
- d) delivery problem in cities

The authors divided (b) and (g) like separated problem, because usually it is necessary to use the special approach to solve it in a real life.

Delivery problem in cities (DPC) description is following. It is necessary to deliver the particular amount of goods Q through the known-before route with n roads' segments, providing particular m objects with the needed d_j units of cargo.

$$\sum_{j=1}^m d_j = Q \quad (4)$$

The most important is the restrictions of the model: each road segment distance is known before, but the speed of a vehicle in each road segment i may be different depending on the day of the week or hour of a day as well as amount of cargo q_i may change for objects m ; unloading processes labour-intensity also changes as a result of the different objective circumstances.

Of course, there are many known methods and algorithms to optimize circular route planning and reduce delivery costs as well as vehicle's run and transportation costs, but these usage often is not efficient to plan goods deliveries in cities by optimal way due to traffic intensity fluctuation by hour of a day or day of a week. In additional, often it is impossible to provide the needed result, because use of these methods is connected with the following problems: using the method is work-consuming or it is impossible to provide neither accurate result, nor satisfy customers' individual needs, nor take into account traffic intensity changes depending on a day of the week and hour of a day.

The authors conclude that computer programs often provide non-optimal solution to transport problem due to different restrictions; on the one hand, it is possible to use only these programs to solve theoretical problems. On the other hand the real situation is changing daily, because demand is not stable, it is possible to use heuristic methods. But in a real life computer programmes are used in combination with heuristic method to achieve the optimal result. Because of lack of information, criteria of optimization used in practical conditions often are vehicle's run and transportation costs, not delivery time factor.

Traffic intensity uncertainty makes inefficient a usage of traditional route planning and optimization methods (for cities), based on vehicle's run or transportation costs minimization, assuming that vehicle's speed is fixed or constant. Therefore, it is necessary to use special approaches and new technologies, to solve delivery problem in cities in the real life.

After that managers may plan circular routes to serve a couple of customers within one route. Planning of the optimal customers' serving order allows improving of small shipments delivery system especially in cities with hard traffic. Problems like VRP (The Vehicle Routing Problem), TSP (Travelling Salesman Problem), SRP (Street Routing Problem) also are known for a long time and often require individual solution for the particular situation. It is expedient to use Minimal Growth Method (MGM) to plan optimal customers' serving order within one circular route. This help to improve customers' serving quality system, minimizing total vehicle's run or/and delivery time, or/and delivery cost. Thus, transport managers may use MGM to plan competitive deliveries in cities and other built-up areas.

SOLUTION WITH THE MINIMAL-GROWTH METHOD COMBINATION EXAMPLE

The authors worked out three steps' algorithm for small shipment delivery's quality improvement in cities with unstable traffic (figure 3). It is expedient to make following steps to improve quality of small-shipment deliveries in cities and define the optimal customers' serving order:

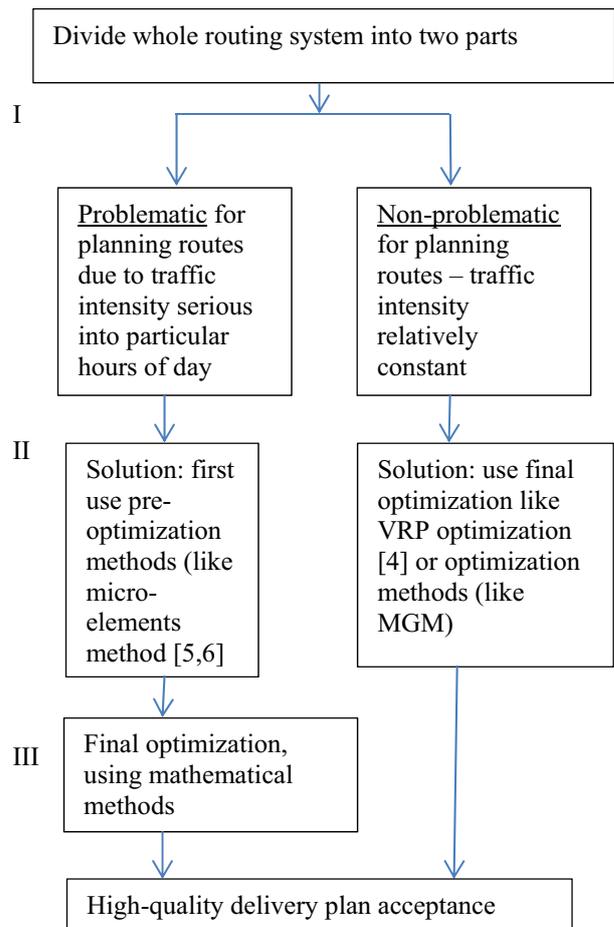


Figure 3. Three steps for small shipment delivery's quality improvement in cities with unstable traffic.

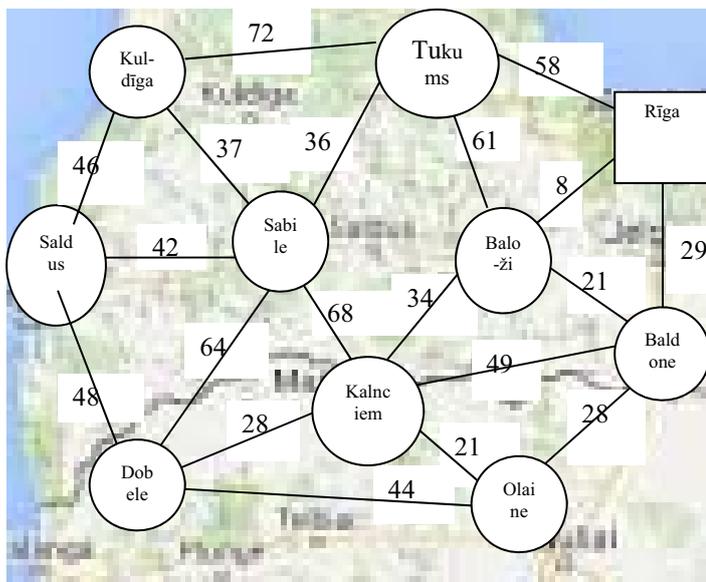
Actually,, it is possible to use only micro-elements method [7,10] like pre-optimization methodology to improve problematic route planning in cities with intensive and traffic (not analyzed into this paper in details; the problem analyzed into the other P.Patlins's papers[6,7,10]). Specialists divide vehicle's moving time into separated elements to make easy and more accurate vehicle's moving time. Therefore, it is expedient to use minimal-growth method to plan circular routes in cities with hard traffic.

There are some ways how to improve non-problematic routes planning in cities and towns. Necessary to find a solution of Travelling Salesman problem or Vehicle Route problem to serve couple of customers, grouping them within one route, which starts and finishes at the same warehouse to minimize total vehicle's run or/and delivery time, or/and delivery cost, when given following information: Customers coordinates; each customer's demand q_i ; distance between each two customers: l_{ij} ; delivery cost of each route segment: C_{ij} ;driving time between each two customers: t_{ij} ; capacity of vehicle is Q_i ;

The model allow to calculate *the optimal quantity of vehicles* to serve customers as well as *customers ' serving order* for each route to reduce vehicle's run or/and delivery time, or/and delivery cost:

$$\sum_{i,j=1}^n l_{ij} \text{ or } \sum_{i,j=1}^n C_{ij} \text{ or } \sum_{i,j=1}^n t_{ij} \rightarrow \min \quad (5)$$

For example, company has warehouse (Riga) and some customers (other cities). It is necessary to serve customers by optimal way, finding the best order to serve them (Figure 2)- see customers location scheme.



Riga – warehouse/wholesaler/logistics company
Tukums, Kuldīga, Sabīle, Saldus, Dobele, Kalnciems, Baloži, Baldone, Olaine – customers/serving objects

Figure 4. Location of the customers and warehouse.

Customers' demand given in the table 1.

Table 1. Customers' demand

Customer	Demand, kg
Dobele	375
Saldus	500
Kalnciems	400
Balozi	425
Baldone	500
Tukums	575
Sabīle	125
Olaine	675
Kuldīga	425

Vehicle's capacity is 2 000 kg

How to find the optimal quantity of vehicles to serve customers as well as customers ' serving order for each route to reduce vehicle's run ?

First of all it necessary to build so called "minimal three" or the shortest objects' connection network to connect all objects of the task, using the nearest neighbour method.

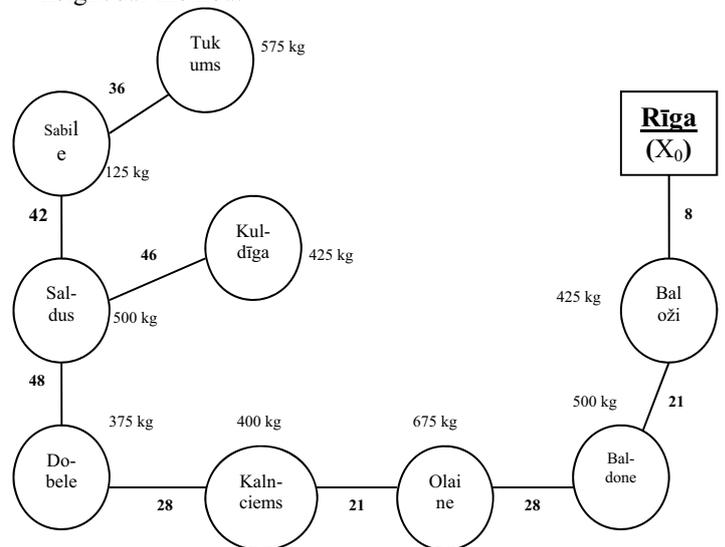


Figure 5. The "minimal three"

Checking objects' demand as well as vehicles' capacity may conclude, that it is necessary to plan two routes (see table 2).

Table 2. Customers division into two routes

Route I		Route II	
Objects	Demand, kg	Objects	Demand, (kg)
Tukums	575	Kalnciems	400
Sabile	125	Olaine	675
Saldus	500	Baldone	500
Kuldīga	425	Baloži	425
Dobele	375		
Sum:	2000	Sum:	2000

The next step is: to find the best customers serving order for the first route. First of all build the shortest ways matrix for the first route (table 3), using information about objects' location from the Figure 2. Calculate the total sum for each column of the matrix.

Table 3. The shortest ways matrix for the first route

Rīga (R)	58	94	103	130	62
58	Tukums (T)	36	52	72	39
94	36	Sabile (Sb)	42	37	64
130	72	37	Saldus (Sl)	46	89
103	52	42	46	Kuldīga (K)	48
62	39	64	48	89	Dobele (D)
$\Sigma 447$	$\Sigma 257$	$\Sigma 273$	$\Sigma 291$	$\Sigma 374$	$\Sigma 302$

After that choose 3 biggest sums [4] from the last line of the table 3 as well as appropriated objects: A, H and B.

Create a basis of the route.

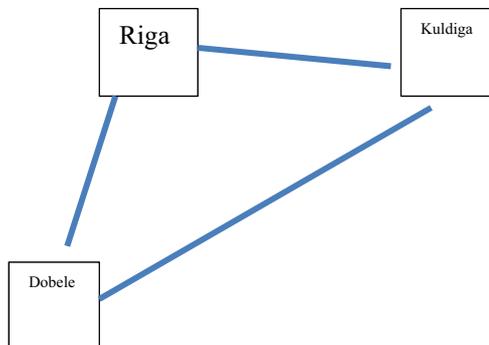


Figure 6. The basis of the route.

The next step is: to choose the column with the next biggest sum and include appropriated point in the

route, using following principle:

$$\Delta = C_{fn} + C_{ns} - C_{fs}, \quad (6)$$

where:

- growth of the route, putting new object into appropriated routes' interval. .

C - distance, km;

n- new customer's index; ;

f- first point of the pair;

s - second point of the pair.

So, calculate Δ for G:

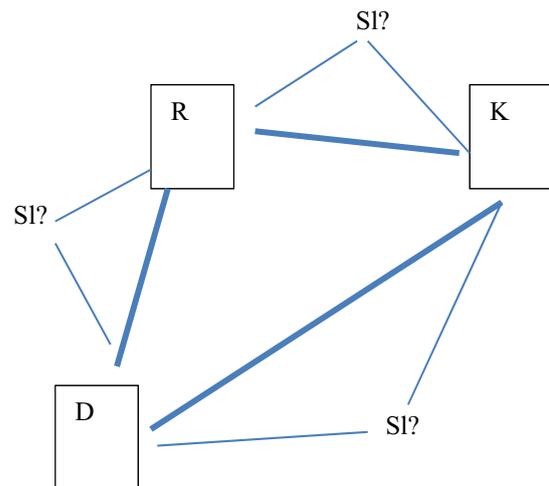


Figure 7. The interval for G customer

Using formula 6:

$$\Delta_{RK} = C_{RSI} + C_{SIK} - C_{RK}$$

$$\Delta_{KD} = C_{KSI} + C_{SID} - C_{KD}$$

$$\Delta_{DR} = C_{DSI} + C_{SIR} - C_{DR}$$

So:

$$\Delta_{RK} = 103 + 46 - 130 = 19 \text{ km}$$

$$\Delta_{KD} = 46 + 48 - 89 = 5 \text{ km} - \text{The minimal growth.}$$

Put SI between K and D.

$$\Delta_{DR} = 48 + 103 - 62 = 89 \text{ km}$$

Repeat the same steps for Customer Sb :

$$\Delta_{RK} = C_{RSb} + C_{SbK} - C_{RK} = 94 + 37 - 130 = 1 \text{ km} - \text{the minimal growth. Put Sb between R and K.}$$

$$\Delta_{KSaI} = C_{KSb} + C_{SbSI} - C_{KSI} = 37 + 42 - 46 = 33 \text{ km}$$

$$\Delta_{SaID} = C_{SISb} + C_{SbD} - C_{SID} = 42 + 64 - 48 = 58 \text{ km}$$

$$\Delta_{DR} = C_{DSb} + C_{SbR} - C_{DR} = 64 + 94 - 62 = 96 \text{ km}$$

Then repeat the same steps for customer E:

$$\Delta_{RSb} = C_{RT} + C_{TSab} - C_{RSb} = 58 + 36 - 94 = 0 \text{ km}$$

Finally plan the optimal route A - H - G - E - C - B - A (Figure 8) which provides the shortest way to serve all customers and return to the company' warehouse.

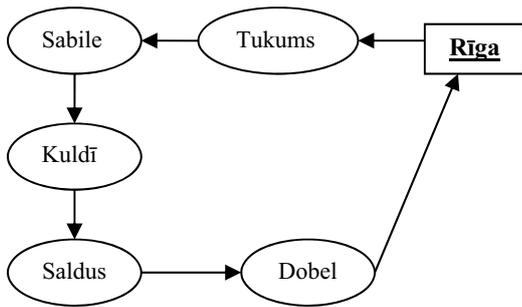


Figure 8. Route I. The optimal customers' serving order.

Repeat the algorithm to find the optimal solution also for the second route (Figure 9).

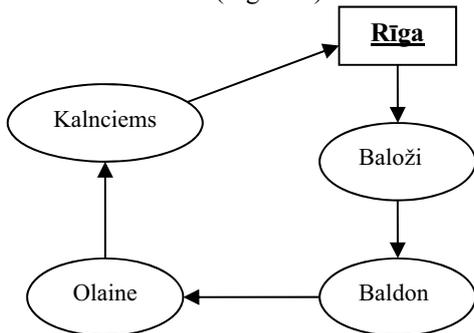


Figure 9. Route II. The optimal customers' serving order.

The method provide the best results for “non-problematic” circular routes, planning small- shipment deliveries in cities and other built-up areas. The Minimal Growth Method is quite simple and effective to be used into practical situations.

CONCLUSION

Various computer programs often provide non-optimal solution to transport problem due to different restrictions; on the one hand, it is possible to use only these programs to solve theoretical problems.

Usage of minimal-growth method allows planning the optimal serving order of customers within the route as well as minimizing of the total transportation costs, time or vehicles's run.

The minimal-growth method is an universal method for circular routes planning; it allows to solve a lot of problems and serve customers by optimal way. The method allows also to divide a couple of customers into particular groups and “connect” them with the vehicle depending on its capacity and define needed quantity of vehicles and routes for the best planning.

REFERENCES

1. Gadzhinski, A.M. Logistics, 3rd edition. Moscow, 2000.
2. Jönson, G and Johnsson, M. Packaging logistics in product development, 5th International Conference on

Computer Integrated Manufacturing , ICCIM 2000.

2. Klanšek, U, Pšunder M. Solving the nonlinear transportation problem by global optimization. Transport: Volume 26, Issue 4, Vilnius Gediminas Technical University, Vilnius 2011.

3. Matis, P. Decision support system for solving the street routing problem. Transport: Volume 23, Issue 3, Vilnius Gediminas Technical University, Vilnius 2008, 320-235.

4. Nerush, J.M. Logistics, 2nd edition. Moscow: Junity, 2001.

5. Patlins, P. Local deliveries time optimization for cities with unstable traffic. Proceedings: 22nd European Conference on Modelling and Simulation. Cyprus, Nicosia, 2008.

6. Patlins P. Circular routes planning improvement for cities with intensive traffic. Proceedings of the International Conference of Logistics, Informatics and Service Science. Beijing Jiaotong University, China, June 2011.

7. Patlins P. Logistics service Quality Improvement for City Routes // Proceedings of the International Conference of Industrial Logistics. 14-16 June 2012 . Zadar, Croatia. ISBN:978-953-7738-16-7

8. Prins, C. A simple and effective evolutionary algorithm for the street routing problem. Computers & Operations Research, Volume 31, Issue 12, October 2004, 185-202.

9. Sprancmanis N, Patlins P. Preču fiziskās sadales organizēšana intensīvas satiksmes apstākļos. RTU Zinātniskie raksti, 3.sērija, 12. sējums. Rīga, 2006, 144-149.

11. Tolley, R. and Turton, B. Transport systems, Policy and Planning. A Geographical Approach. Longman: Harlow,- 1997, 34-41.

AUTHORS' BIOGRAPHIES

Pavels Patlins was born in Riga, Latvia and studied Economics and Logistics and obtained his Logistics Doctor degree in 2011 in Riga Technical University. He has been working for the Consulting companies as a logistics expert as well as for Riga Technical University since 2003, conducting Logistics and Transport Managements lectures. His e-mail is: pavels.patlins@rtu.lv.

Remigijs Pocs. is a dean of the Faculty of Engineering Economics and Management; Head of Division of Foreign Economic Relations, Transport Economics and Logistics at Riga Technical University. He is an Expert of the Latvian Council of Science; Head of Promotion Council (Economics), Riga Technical University; Member of the Council of Professors of RTU Faculty of Engineering Economics and Management, LU Faculty of Economics and Management and LUA Faculty of Economics; Scientific Editor -in-Chief, Scientific Proceedings of Riga Technical University “Economics and Entrepreneurship”. His e-mail is: remigijs.pocs@rtu.lv

TRACKING BUSINESS TRENDS – DILEMMAS OF MEASUREMENT

Péter Juhász
János Száz
Kata Váradi
Ágnes Vidovics-Dancs

Department of Finance
Corvinus University of Budapest
H-1093, Budapest, Hungary
E-mail: kata.varadi@uni-corvinus.hu

KEYWORDS

Trends, fluctuation, corporate performance, inflation, leverage

ABSTRACT

Corporate performance may be tracked using various measures. Our model simulating the behaviour of a simple firm underlines that the choice on measurement unit determines what distortions we will face so based on different measures we may end up identifying completely different cycles. On the top of that, these cycles would radically modify if firms within the given industry would change their strategy towards the same direction or when structural changes happen in the economy. This may end in researchers analyzing non-existing cycle changes.

INTRODUCTION

We may measure corporate performance in various ways. Total sales, operational profit, or after tax profit are used by market analysts to describe a given industry, sum of added value (GDP – gross domestic product) is common measure in macro papers, while at firm level owners may focus on dividends, cash flows, or some profitability ratios, like ROI (return on investment), ROE (return on equity), or CFROE (cash flow return on equity). We may assume that an industry of well performing companies should be doing well at sector level, and an economy consisting of boosting industries ends up with great trends in macro economy. This argumentation may be logical, but is that really true once we use different measures to access performance at each of those levels? Our model shows how measurement results may differ across measures in case of a simple company when controlling for (1) operational and (2) financial leverage, (3) equipment lifetime, (4) demand fluctuations, and (5) inflation.

The main goal of our research is to show how the performance can differ depending on what level we carry out our analysis: on the whole economy level, the industry level or the company level. This is an important question, since in the literature several papers are dealing with the question of performance

measurement, but usually the researches focus on one of the levels, and on different indicators, and different analysis methods. For example on the whole economy level related to performance, a key issue is to handle business cycles. The research related to the measurement of business cycles goes back to the late forties. The first notable research was carried out by Burns and Mitchell (1946). This research was followed by several more in this field. The main focus of these research were how to decompose the business cycle component from the empirical datasets, e.g. Baxter and King (1999), Hodrick and Prescott (1997), Darvas and Szapáry (2004), Hassler et al. (1992) or Diebold and Rudebusch (1994). The most commonly used methods based on Baxter and King (1999) are the following: two sided moving average; first-differencing; removal of linear or quadratic trend; application of Hodrick-Prescott (1997) filter; and band-pass filter. The variables that the researches usually use to analyze business cycles, are usually some type of macroeconomic factor, such as GNP (gross national product), fixed investment, employment, etc.

While on industry and company level the performance is measured in various way also empirically and theoretically. For example Capon et al. (1990) are using a meta-analysis method to analyze corporate performance by applying financial and non-financial indicators. They collected the indicators based on the empirical literature of the industry and company level based researches between 1921 and 1987.

Since in our paper we will focus only on financial indicators, we will use those ones, which are usually used in the literature, like the Sales, EBIT or the ROE (Damodaran, 2012).

Besides relying on the financial indicators generally used in the literature, we will also take into account the results of the literature in another aspect as well. Since in previous researches it was found that the effect of inflation is notable regarding the profitability and the value of a company (Dömötör et al., 2013, Radó, 2007). According to this we will use inflation in our models.

The paper will be built up as following: first we will introduce our model, then we show our results in a base scenario, where the demand on the market is stable, and the leverage of the companies is zero as

well. Then in the following chapters we show how the fluctuation in demand will effect the performance of the company, the industry, and the whole economy. We will also analyze the effect of the different operational and financial leverages, the inflation, and the lifespan of the equipments to the performance. Finally in the last chapter we have our conclusions, and the limitations of our research.

MODEL DESCRIPTION

Our model tracks the performance of one single simplified firm. The company has only one product, which is manufactured using one type of machine. The net working capital of the operation is zero – payables financing inventory and customers completely –, so invested capital (IC) equals to the total value of the equipment.

The sales price (10) and demand quantity (2000 in the first period) is determined by the market forces and cannot be influenced by the firm itself. At the same time, the management will have an exact prediction of the demand at the beginning of each period, so they can purchase exactly the needed amount of machines and will manufacture all products that the market asks for. Though, they may not sell equipment purchased in the previous periods. Capacity only decreases once lifetime of the machine is over.

The firm has variable costs depending on the quantity produced and fixed costs that do not change with the amount produced. A pre-set part (50%) of both cost types is labour expense. Both sales and all types of manufacturing costs grow at the same inflation rate. To allow for comparison we set manufacturing costs always so that during the first period the firm earns an operational profit before depreciation and amortization (EBITDA – earnings before interest taxes depreciation and amortization) of 8000.

There are several kinds of machines available for the production all able to produce the same amount (10 thousand pieces) of product during a period. Those only differ in their useful lifetime (from 1 up to 6 years) and are depreciated linearly. The cost of each machine is calculated so that the yearly cost equivalent for each type would be the same. The price of the machine is indexed to inflation across periods and only whole number of machines can be bought.

At the start of period 1 we always assume that the machines owned are just enough to serve the first period demand and had been purchased in equal quantities during the previous years, so those will need gradual replacement. Given the different lifespans of the equipment when the required product quantity on the market changes the company may have to purchase new machines earlier or accumulate unused capacity depending on the type of machine used.

To calculate operational profit (EBIT – earnings before interest and taxes) manufacturing costs and D&A (depreciation and amortization) is deducted from sales. Then cost of debt (interest) is accounted for, and corporate tax (20%) is deducted to calculate profit after

tax (PAT). The interest rate is automatically indexed for inflation. Retained earnings is calculated based on the required growth of equity given product demand of the next year. The difference of PAT and retained earnings is the sum of dividend paid and equity raised or repurchased. This is the cash flow that owners will face (FCFE – free cash flow to equity) and which would determine in the real life the market value of the ownership.

BASE SCENARIO

In the base scenario there is no growth or fluctuation in market demand, no inflation, and we have variable manufacturing costs only (operational leverage=0), operate without debt (financial leverage=0). Due to this, all periods modelled look the same.

Depending on the management choice of machines (financially completely value neutral) we will see different investment need, D&A, EBIT, tax, PAT, and dividend (FCFE). Though, sales and added value (AV = EBIT + D&A + Labour expenses) are the same in any case. As the choice of machine influences the investment need (IC), ROI, ROE and CFROE differ also heavily.

Table 1. Comparison across machine types

Level of analysis	Performance measure	Lifetime of machines		
		1 year	3 years	6 years
Macro	Added value	14 000.00	14 000.00	14 000.00
Industry	Sales	20 000.00	20 000.00	20 000.00
Industry	EBIT	2 990.38	3 409.16	3 920.00
Industry	PAT	2 392.31	2 727.33	3 136.00
Firm	ROI	59.69%	37.13%	27.45%
Firm	ROE	47.75%	29.70%	21.96%

Table 1 illustrates the differences between firms using machines of 1, 3 and 6 years of useful lifetime. We may conclude that while macro analysts would see no difference between the firms, industry analyst would see better performance at firms with machines of longer useful lifetime. At the same time, owners of the firms with shorter lifetime assets would be happier due to higher returns achieved.

Operational leverage would have no effect here, as costs are not changing over time, while financial leverage decreases PAT and boosts ROE (once cost of debt is less than ROI). The effect of inflation may seem neutral for the first look, as both sales price and all types of expenses are inflated by the same percentage. This is indeed true for Sales, Added value, and investment but not for EBIT, PAT, and FCFE (dividend) once the useful lifetime of the machines is longer than 1 year as it is shown in Figure 1.

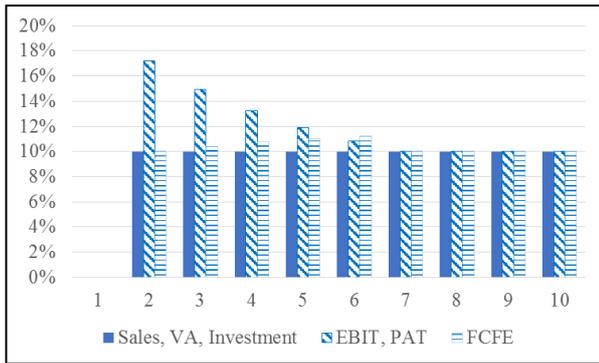


Figure 1. Yearly growth rate at 10% inflation for machines of 6 years lifespan

The reason for this is that D&A is not indexed for inflation, so it takes time that it reflects the growing price level. The lower than realistic D&A increases EBIT and PBT (profit before tax). As PBT is increased so by more than the inflation rate, the real tax burden of the companies grows. As invested capital (and so equity) is not indexed by inflation either, ROI and ROE grow also radically. This phenomenon is also illustrated by Table 2 and 3.

Table 2. Effect of inflation on the first year's numbers (1)

Machine lifetime	1 year			
	Inflation	0%	10%	Change
Added value		14 000.00	15 400.00	10.00%
Sales		20 000.00	22 000.00	10.00%
EBIT		2 990.38	3 790.38	26.75%
Tax		598.08	758.08	26.75%
PAT		2 392.31	3 032.31	26.75%
ROI		59.69%	68.78%	15.23%
ROE		47.75%	55.03%	15.23%
IC		5 009.62	5 009.62	0.00%
E		5 009.62	5 009.62	0.00%

Table 3. Effect of inflation on the first year's numbers (2)

Machine lifetime	6 years			
	Inflation	0%	10%	Change
Added value		14 000.00	15 400.00	10.00%
Sales		20 000.00	22 000.00	10.00%
EBIT		3 920.00	4 720.00	20.41%
Tax		784.00	944.00	20.41%
PAT		3 136.00	3 776.00	20.41%
ROI		27.45%	32.14%	17.06%
ROE		21.96%	25.71%	17.06%
IC		14 280.00	14 280.00	0.00%
E		14 280.00	14 280.00	0.00%

This means that depending on the average useful lifetime of machines applied a suddenly appearing

inflation may distort statements for several years showing improvement in some of the measures while leaving other unchanged. On the top of all that the exact extend of distortions is also dependent on the type of equipment used by the firm.

INTRODUCING DEMAND FLUCTUATION

To get a more realistic model we assume some fluctuation in demand overtime according to Equation 1. To keep it simple we use a sinus function to achieve cycles between 2 and 3 million pieces per period. Figure 2 and 3 contrast the development of key quantities in case of different machine types. Our equation for demand (Q) is as follows:

$$Q_t = Q_0 + a * (1 + \sin(c * t)) \quad (1)$$

For the sake of example a=500 and c=100 have been chosen as parameter values.

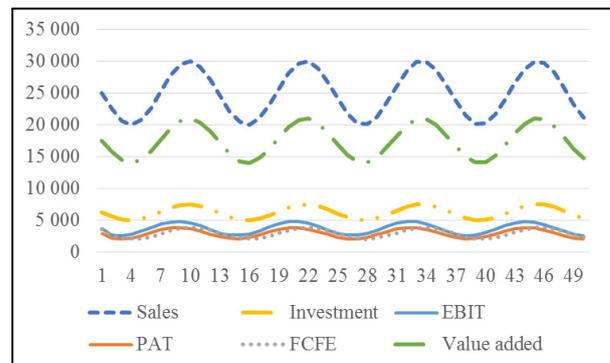


Figure 2: Effect of demand fluctuation – Lifetime of machines: 1 year

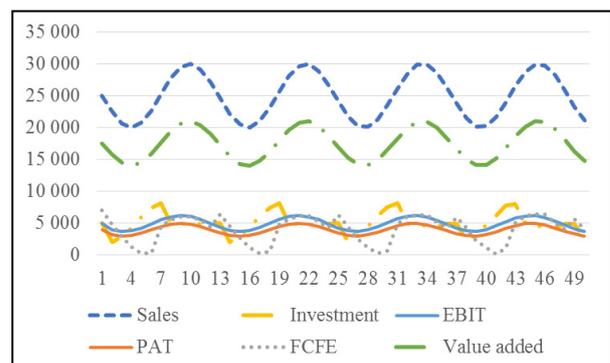


Figure 3: Effect of demand fluctuation – Lifetime of machines: 6 years

Note, that longer lifetime leads to investment and FCFE following new patterns. It is key to see that even during times of increasing output, Sales and added value investment may fall back as current capacity is dependent not only on current investment level, but also on those of the previous 5 years. Due to this fluctuation FCFE may not only grow when performance increases, but also when lower proportion

of current profit is needed to keep production capacity at the required level.

Differences are more dramatic when focusing on financial ratios instead of absolute quantities. As Figure 4 shows that the previously experienced synchrony disappears: in case of using 1-year machines CFROE, ROI, and ROE are unchanged and equal as the firm can adapt to the market fluctuations perfectly. When using equipment with 6-years life time, company will have some unused capacity during some periods, destroying capital efficiency. This means, that the risk of shares will also differ.

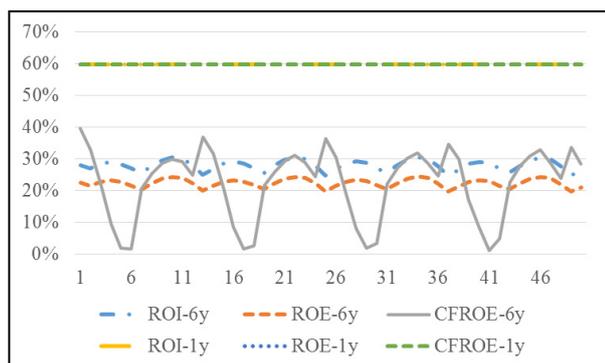


Figure 4: Effect of demand fluctuation – Financial ratios

FLUCTUATION AND LEVERAGE

Now, that manufactured amount changes from period to period, the amount of operational leverage (percentage of fixed costs) plays an important role. Assume that two technologies exist: the one used until now with 6 units of variable cost (VC) per piece and no fixed costs (FC), and another with 4 units of VC and 5000 units of FC. Note, that both of these technologies imply an EBIT of 4960.

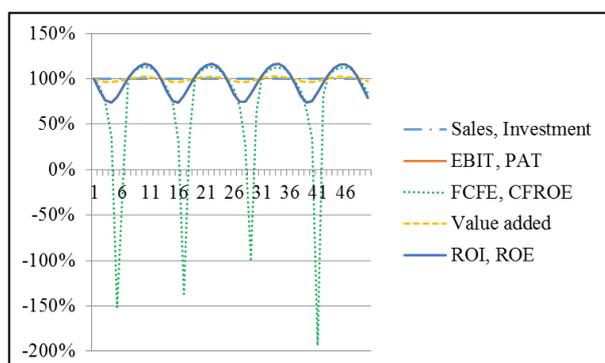


Figure 5: Performance with operational leverage in percentage of that without leverage (machines used for 6 years)

The only two measures that operating leverage does not affect is sales and investment. As fixed cost do not change overtime, more fluctuation is to be seen in all other quantities. Figure 5 offers a comparison between

two otherwise identical firms using two different technologies.

Financial leverage (we assumed $D/IC=50\%$, interest=10%) only effects P/L (profit and loss) item below EBIT. PAT is lowered by interest payment, but only hurts FCFE in periods when ROI is lower than cost of debt. In all other periods FCFE is dramatically increased that results in boosted CFROE at any time due to the continuously lower equity requirement as it can be seen in Figure 6. It is also worth noticing that operational leverage increased risk by boosting downside potential, while financial leverage (under the given conditions) added to risk by letting the upside grow.

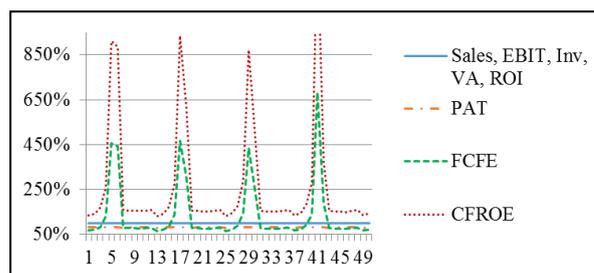


Figure 6: Performance with financial leverage in percentage of that without leverage (machines used for 6 years)

Adding inflation to the fluctuations will also complicate trend analysis. The steady price growth pushes up profits faster than sales or AV due to the lagging historical prices in D&A. As book value of machines (IC) is not indexed by inflation while profit is higher due to the D&A effect. ROI and ROE distortedly shows a better performance. CFROE is more realistic as D&A effect is not hitting it. FCFE shows radical fluctuations as the demand fluctuation requires to buy a huge number of new equipment every twelfth year but as FCFE is growing slower than investment, those years equity needs to be raised to cover extra investment, while the real performance of the firm has not changed at all.

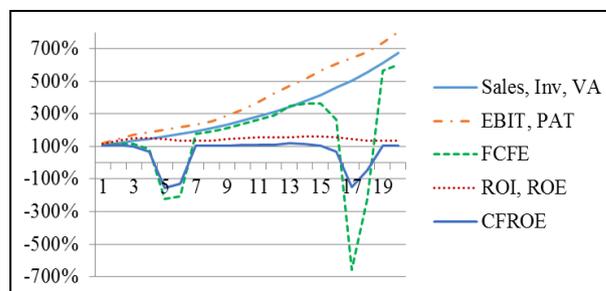


Figure 7: Performance with inflation (10%) in percentage of that without inflation (machines used for 6 years)

As we have seen, once demand is not constant it is not only the useful lifetime of the equipment used but also

operational and financial leverage, and inflation that would modify the measurable performance trends. In the next step we investigate how all these factors together may influence the financial numbers of a firm. Let us compare the development performance measures of two firms facing the same demand trends but using different machines (1 year lifetime against 6 years life time), different technology (VC=6 only and VC=4 and FC=5000), and different financing (D/IC=0 and D/IC=50% interest=10%). For simplicity we assume these firms operate in the same country and face the same inflation (0%). Note that the first firm is identical to what appears on Figure 2.

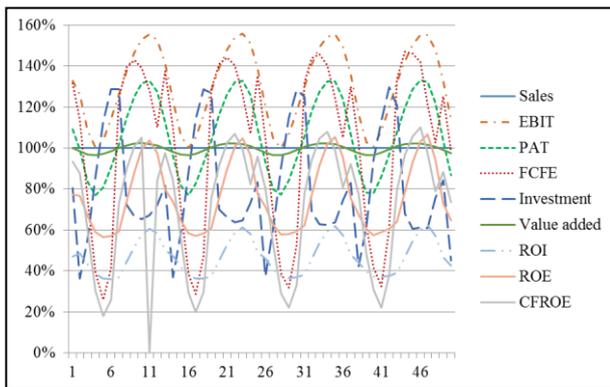


Figure 8: Comparing firms from the same industry of different equipment, technology, and financing ($a=500$, $c=100$) (ratio of performance measures)

Figure 8 illustrates the performance measurement problem of a given sector. Though sales trends are just the same (flat line at 100%), all other performance measures would differ across firms due to individual characteristics. It is easy to see that distortions are very different both in size, form, and timing. So when aggregating (summing, averaging) certain performance measures we would end up concluding totally different trends for the whole industry altogether.

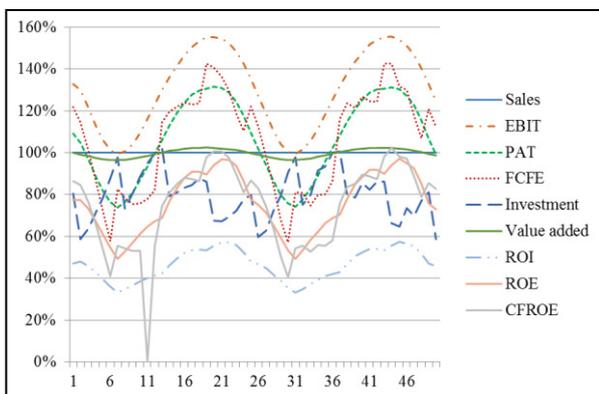


Figure 9: Comparing firms from the same industry of different equipment, technology and financing ($a=500$, $c=50$) (ratio of performance measures)

While one might think that careful modelling may help us to get rid of these distortions. Figure 9 and 10 supports that the problem is more complex. Just by increasing the wave length of the demand fluctuation to its twofold or fourfold (slower fluctuation of the same size) leads to a very different set of differences. Distortions in performance measures become more similar as wave length increases. (Endlessly long waves can be very similar to the flat demand we used at the beginning of this paper.)

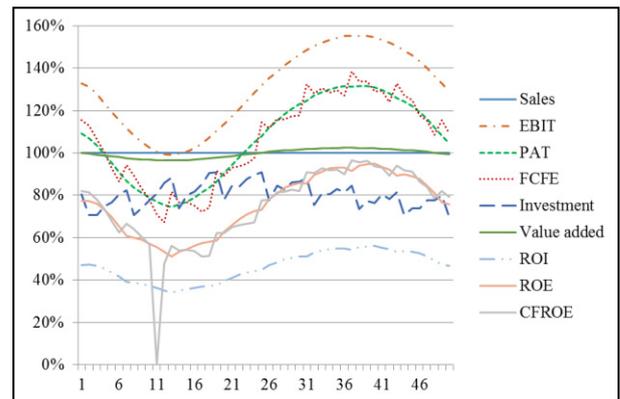


Figure 10: Comparing firms from the same industry of different equipment, technology and financing ($a=500$, $c=25$) (ratio of performance measures)

It is important to notice that while in case of the original fluctuation (Figure 8) ROE was able to over perform at peak times of the base model (Figure 2), due to change in the wave length this was not possible anymore. In other words it is also the type of demand fluctuation that determines how successful a given strategy might be on the market.

CONCLUSION

We prepared a simplified financial model of a manufacturing firm and analyzed how the useful lifespan of equipment used (length of replacement cycle), operational and financial leverage are applied (business strategy), and demand fluctuation and inflation (market conditions) would influence performance measures.

Even in case of stable demand the kind of assets used had serious effect on financial performance even though financially all of the alternatives cost the same (equal yearly cost equivalent) – a result quite counterintuitive.

We also saw that appearance of inflation not only increases tax payment in real terms cutting back on the value of the firm but at the same time distorts performance measures to show a contrary trend.

When demand fluctuation was introduced into the model it has become clear that investment may not peak in periods where demand does depending on the length of equipment lifetime and due to that cash flow

to shareholder may also be higher in years with lower demand.

The use of short lifetime equipment seems to protect owners from fluctuations of profitability ratios, while operational and financial leverage increase risk. Though added operational risk shows in increased downside potential while financial leverage (under our assumptions) offered an enhanced upside potential.

When considering inflation a new serious problem was identified: because of demand fluctuations owners were forced to regularly pay in cash to maintain an operation that did not change at all in real terms.

Finally, we compared performance measures of firms with different strategy to figure out that the choice of firms on machines and leverage would have dramatic effect on the performance measures making the original demand trend nearly unrecognizable. Depending on what kind of measure we focus on, the industry cycle would be described completely differently.

This issue becomes particularly important in a transforming economies. Once companies tend to change their strategy (some technologies, machine types, particular leverage level gaining popularity) or the structure of economy is shifted preferring firms with a given strategy, we may measure macro trend changes that are not existing at all.

Unfortunately these distortions are not even stable but rather depend on the speed of market fluctuations. It is not only the size but also the speed of market fluctuations that determine how successful a business strategy would be.

Due to these we have to be very careful when choosing a metric to track financial performance across time of a given industry of firm. Even an unchanged strategy could lead to very wild fluctuations in performance on a relatively stable market when different waves interpolate.

LIMITATIONS

In the real life firms may not be able to precisely predict the quantity to be produced and sold during the next period that may lead to distortions in investments and manufacturing. No matter whether they over or under estimate demand they will have a worse performance than predicted by our model, as both unneeded capacity and market growth potential not completely used causes losses compared to the optimum.

Inflation rates may differ across various types of cost, particularly the increase of wages may be very different to that of the material expenses. This could lead to even more complex distortions.

REFERENCES

Baxter, M. and King, R.G., 1999. Measuring business cycles: approximate band-pass filters for economic time series. *Review of economics and statistics*, 81(4), pp.575-593.

Burns, A.F. and Mitchell, W.C., 1946. Measuring business cycles. *NBER Books*.

Capon, N., Farley, J.U. and Hoenig, S., 1990. Determinants of financial performance: a meta-analysis. *Management Science*, 36(10), pp.1143-1159.

Damodaran, A., 2012. *Investment valuation: Tools and techniques for determining the value of any asset* (Vol. 666). John Wiley & Sons.

Darvas, Z. and Szapáry, G., 2004. *Business Cycle Synchronization in the Enlarged EU: Comovements in the (soon-to-be) New and Old Members* (No. 2004-1). Magyar Nemzeti Bank Working Paper.

Diebold, F.X. and Rudebusch, G.D., 1994. Measuring business cycles: A modern perspective (No. w4643). *National Bureau of Economic Research*.

Dömötör, B., Juhász, P. and Száz, J., 2013. Devizaárfolyamkockázat, kamatláb kockázat, vállalatfinanszírozás (Foreign exchange risk, interest rate risk, financing). *Hitelintézet* Szemle. Vol. 12. No. 1. pp. 38-55.

Hassler, J., Lundvik, P., Persson, T. and Söderlind, P., 1992. *The Swedish business cycle: Stylized facts over 130 years* (No. 22). Institute for International Economic Studies, Stockholm University.

Hodrick, R.J. and Prescott, E.C., 1997. Postwar US business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pp.1-16.

Radó, M., 2007. *Az infláció hatása a vállalati értékkülönös tekintettel az adóhatásokra* (Doctoral dissertation, Budapesti Corvinus Egyetem).

AUTHOR BIOGRAPHIES

PÉTER JUHÁSZ is an Associate Professor of the Department of Finance at Corvinus University of Budapest (CUB). He holds a PhD from CUB and his research topics include business valuation, financial modelling, and performance analysis. His e-mail address is: peter.juhasz@uni-corvinus.hu

JÁNOS SZÁZ is a full Professor at the Department of Finance at Corvinus University of Budapest. He is the first academic director 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

KATA VÁRADI is 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 areas are market liquidity, bonds markets and capital structure of companies. Her e-mail address is: kata.varadi@uni-corvinus.hu

ÁGNES VIDOVICS-DANCS is an Assistant Professor of the Department of Finance at Corvinus University of Budapest. Her research topics 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

Simulation of Intelligent Systems

FAST 3D HOUGH TRANSFORM COMPUTATION

Egor. I. Ershov, Arseniy P. Terekhin
Simon M. Karpenko, Dmitry P. Nikolaev
Institute for Information
Transmission Problems, RAS
127994, 19, Bolshoy Karetny per.,
Moscow, Russia
E-mail: ershov,ars,simon,dimonstr@iitp.com

Vassili V. Postnikov
JSC Cognitive,
117312, office 709, 9, pr. 60-letiya Oktyabrya,
Moscow, Russia
E-mail: vassili.postnikov@gmail.com

KEYWORDS

Fast Hough transform, Dyadic planes, Exhaustive search, Radon transform, Fast algorithms, Image processing

ABSTRACT

We present a three-dimensional generalization of linear Hough transform allowing fast calculating of sums along all planes in discretized space. The main idea of this method is multiple calculation of two-dimensional fast Hough transforms combined with a specific method for plane parametrization. Compared to the direct summation, the method achieves significant acceleration ($O(n^3 \log n)$ vs $O(n^5)$).

INTRODUCTION

Hough Transform (HT) was invented by Paul Hough in 1959 for the analysis of bubble chamber photographs, and patented in 1961. Later HT was modified by R. O. Duda and P. E. Hart to eliminate cases with unbounded transformation space (Hart 2009). There is a widespread opinion that despite algorithm advantages and applicability to different problems Hough transform is too slow with computational complexity being $O(n^3)$.

Fortunately, there is a fast modification of Hough transform - fast Hough transform (FHT), that is not so widely known. Complexity boundary for FHT is $O(n^2 \log n)$ for square image with linear size n , similarly to 2D fast Fourier transform. Moreover, FHT doesn't involve complex arithmetic or even multiplications and could be computed in integer domain.

FHT has a rich reinvention history - we've found four invention precedents. The first one was made in 1995 by W. A. Gotz and H. J. Druckmiller (Gotz and Druckmiller 1995). Further, Martin Brady reinvented FHT in 1998 (Brady 1998), and several years later in 2004 the version with in-place calculations was proposed (Karpenko et al. 2004), and finally, the last reinvention was made by M. Frederick, N. VanderHorn and A. Somani in 2005 (Frederick et al. 2005). Still, newly published HT applicability surveys do not mention FHT, e.g. (Mukhopadhyay and Chaudhuri 2015), (Hassanein et al. 2015).

FHT has become a very popular tool in image processing; a lot of applications of FHT exist, for instance: edge detection, document orientation, vanishing point detection (Nikolaev et al. 2008), detection of circles and ellipses,

linear separation of two-dimensional sets (Ershov et al. 2015a). Also HT was successfully used for robust regression analysis (Ballester 1994; Goldenshluger and Zeevi 2004; Bezmaternykh et al. 2012). We believe, that this tool could improve various computer vision algorithms, e.g. visual odometry and visual localization based on feature point analysis (see Konovalenko et al. (2015); Karpenko et al. (2015)): it allows to use feature lines as well.

Despite of existence and multiple reinventions of FHT for two-dimensional image there is no analogous algorithm for three-dimensional arrays. Such an algorithm could be a very useful tool for many image processing tasks, such as color segmentation, object detection, and orientation estimation using lidar or sonar data, ultrasonic diagnostic, and so on.

In this paper we propose fast three-dimensional Hough transform (3FHT), which calculates sums over all quasi-planes in space using $O(n^3 \log n)$ operations. Here n is linear size of the data cube. We should emphasize that there are two different ways to define 3FHT: as sum along all planes of a cube (discrete Radon transform), or as sum along all lines (discrete Jon transform). From now on we will discuss only discrete analog of Radon transform in three-dimensional space. We use the terms "quasi-plane" or "dyadic plane" to designate discrete planes used in proposed algorithm in contrast to conventional discrete Bresenham planes.

The paper is separated into two chapters. In the first chapter we discuss features of the fast Hough transform for two-dimensional case (2FHT). In the second chapter we describe new 3FHT algorithm, discuss its computational complexity, and geometrical deviation of dyadic plane from its continuous counterpart.

2D FAST HOUGH TRANSFORM

This section is based on materials from (Karpenko et al. 2004) and aimed to emphasize main ideas and features of 2FHT.

Parametrization

Parametrization is one of the main issues while designing Hough transform. A simple form $ax + by + c = 0$ leads to infinite size of Hough space (Hart and Duda 1972). To overcome this problem P. Hart proposed polar parametrization, but unfortunately it doesn't allow to

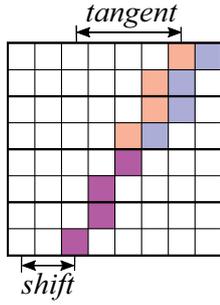


Figure 1: Parametrization and structure of dyadic pattern in two-dimensional fast Hough transform. Two lines with tangent 4 and 5.

construct fast computational scheme, as far as we know. Another parametrization method is shown on figure 1. A line is defined by two positive numbers: shift along one of the rectangle edges s and slope of the line t . Thus all possible lines are divided into four groups: mostly-vertical with right tangent like one on the figure 1, mostly-vertical with left tangent, mostly-horizontal with right tangent, and mostly-horizontal with left tangent. It is easy to see that each group can be transformed to another with reflection or 90° rotation. Thus the computational scheme can be described once. And so the algorithm is described for mostly-vertical lines with right tangent (see fig. 1).

Dyadic Pattern: construction and accuracy

In practice, two Bresenham lines with tangent t and $t+1$ share a lot of pixels or even line segments. Naive HT scheme calculates the same line segment sums multiple times, which leads to computational inefficiency. To overcome this problem Dyadic pattern (structure of discrete line) was proposed. For simplicity we will consider images with linear size $n = 2^p$, where $p \in \mathbb{N}$. To plot dyadic pattern with given tangent D_t one should conduct recursive procedure: at each step the image is divided in half and then initial line segment with slope t is approximated in both halves with shorter line segments having slope $\lfloor t/2 \rfloor$. One pixel shift between these subsegments is added if t is odd. Examples of dyadic patterns are illustrated on the figure 1. Note, that there is no dependence between structure of pattern and *shift*. Mnemonically this rule can be written as

$$D_t = D_{t/2} [t \bmod 2] D_{t/2} \quad (1)$$

We call such type of discrete patterns "dyadic lines" or "dyadic pattern". This construction is recursive in nature. Computation result's reuse allows for complexity reduction from $O(n^3)$ to $O(n^2 \log n)$.

In (Ershov et al. 2015b) it was shown that maximal possible dyadic line deviation from its geometrical counterpart grows with image size as $\frac{1}{6} \log_2 n$. Thus for an image size 1024×1024 the maximal deviation would be less than two pixels, which is good enough for all practical purposes.

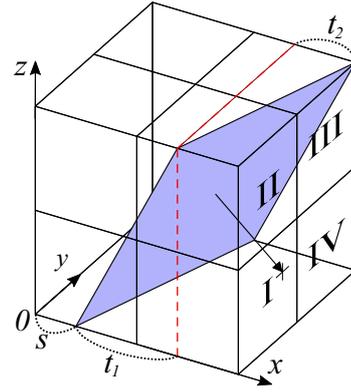


Figure 2: Example of dyadic plane and its parametrization.

3D FAST HOUGH TRANSFORM

In this section we describe new computational scheme for calculating sums along all dyadic planes in data cube. At first glance, it appears to be a huge computational problem. Indeed, the variety of all planes in \mathbb{R}^n is three-dimensional – therefore, one can uniquely represent almost any plane using three parameters a , b and c as in the following equation:

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1 \quad (2)$$

Let us consider planes which intersect given data cube in space. Suppose its edge is of the size n . Then parameters a , b and c would span in $[-2n, 2n]$, so in discrete setting one obtains $12n^3$ planes. It means that the output size of the volumetric Hough transform should be about the same order of magnitude as its input size, which is acceptable. However, for each output cell one seemingly has to calculate the sum over corresponding plane independently. That would take $O(n^2)$ computations per cell, resulting in a huge $O(n^5)$ overall complexity.

Parametrization and Algorithm

Luckily, a considerable speedup is possible. Allowing for modest geometrical inaccuracy, one can compute three dimensional fast Hough transform in $O(n^3 \log n)$, thus spending only $\log n$ summation per output voxel. As far as we know, this result is new. The construction strongly relies on two-dimensional Hough transform described in previous chapter.

Firstly, let us describe convenient plane parametrization. All planes can be divided into twelve groups by normal vector orientation. Indeed, cube has three mutually orthogonal faces, each divided into four parts (see fig. 2). Moreover, normal vector position uniquely defines plane in space. Note that for given normal position it is possible to determine three plane traces. Any pair of which also uniquely determines plane parameters. Therefore, to parametrize the plane, it is enough to fix some point on edge with coordinate $(s, 0, 0)$ and pair of line slopes t_1, t_2 . For simplicity, we will consider further type I planes. It has two mostly-vertical with right slope traces in xy -face (t_1) and in xz -face (t_2). Each trace is right-sloped as illustrated in fig. 2.

Let us sketch out the idea of 3FHT. You can find working MATLAB implementation on github: <https://github.com/Ershoff/FastHoughTransform3D>.

Double integral over any type \mathbb{I} plane that intersects some fixed face F of the cube can be represented as itered integral. First, one have to integrate over all horizontal lines intersecting F . Second, these line integrals should be integrated again over a mostly-vertical lines contained in F . Let's consider this idea in detail.

At the first stage we apply 2FHT to each horizontal xy-slice of the data cube. This way we will obtain another cube (sliced-FHT cube) where voxels represent sum along corresponding lines in horizontal slice. At the second stage we apply 2FHT for each vertical xz-slice of this sliced-FHT cube. Resulting HT-cube contains sum along corresponding plane at each voxel.

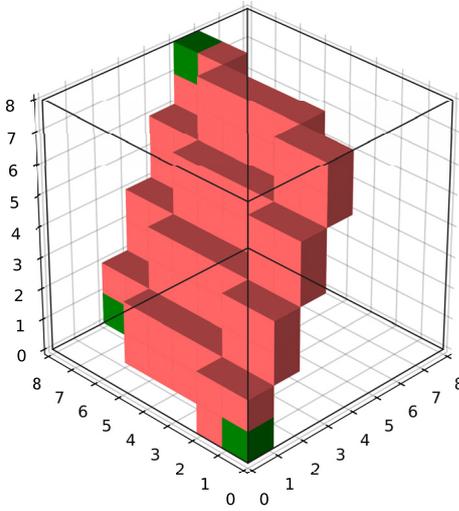


Figure 3: Example of a dyadic plane. This one has $s = 0$, $t_1 = 2$, $t_2 = 4$.

Pseudo-code of FHT3 is described in Algorithm 1. For simplicity, we use function $fht2_quart()$ that performs 2FHT calculation only for the mostly-horizontal lines with right slope. Thus, $fht2_quart()$ takes 2D square array as an input, and returns an array of the same size. To simplify pseudo-code we use “:” notation to work with array dimensions. Operation $dc(:, :, i)$ return a reference to the two-dimensional subarray of dc with $z = i$.

Three coordinates of these voxels correspond to shift s along x-axis and slopes t_1 , t_2 . Example of dyadic plane is shown in figure 3.

Complexity and Precision

Let us consider complexity of proposed algorithm. On the first stage we apply 2FHT to each of n horizontal xy-slices of the data cube. On the second stage we apply 2FHT to each of n vertical xz-slices of the sliced-FHT cube. So we perform two-dimensional fast Hough transform n -times sequentially for both slice types. As 2FHT requires $O(n^2 \log n)$ operations, both stages have $O(n^3 \log n)$ computational complexity. Difference

Algorithm 1 Pseudo-code of 3D Fast Hough Transform

```

function FHT3(dc)                                ▷ dc - data cube
     $n \leftarrow size\_of\_edge(dc);$                 ▷ n - cube edge size
     $hs \leftarrow zeros(n, n, n);$                 ▷ cube filled with zeros.
     $dhs \leftarrow hs;$ 
    for  $i = 1 : n$  do
         $hs(:, :, i) \leftarrow fht2\_quart(m(:, :, i));$ 
    end for
    for  $i = 1 : n$  do
         $dhs(:, i, :) \leftarrow fht2\_quart(hs(:, i, :));$ 
    end for
    return  $dhs;$ 
end function

```

between execution time of naive HT algorithm and 3FHT is illustrated on figure 4.

In previous researches Ershov et al. (2015b), we succeed in proving that maximal spacial distance between corresponding dyadic pattern and ideal line has order $\frac{1}{6} \log_2 n$. Moreover we showed that the largest deviation is achieved in $t = n/3$. It is easy to see that maximal deviation in 3FHT should be twice as big as in 2FHT, i.e. $\frac{1}{3} \log_2 n$.

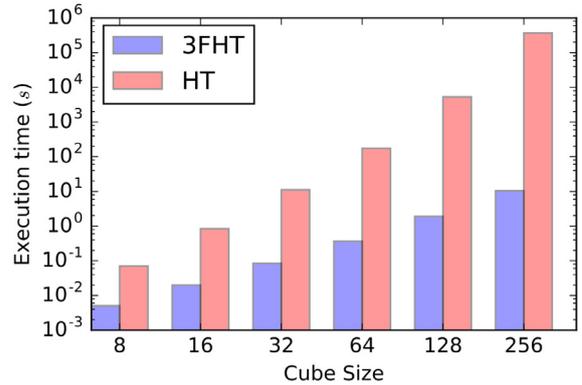


Figure 4: Comparison of computational time for naive HT algorithm and three-dimensional fast Hough transform algorithm. Execution time is given in log scale.

CONCLUSION

In this paper a new effective scheme for calculating three-dimensional Hough transform is presented. Computational complexity of proposed algorithm is reduced from $O(n^5)$ to $O(n^3 \log n)$. To achieve this result we propose novel three-dimensional dyadic pattern and plane parametrization. We state that the maximal deviation of dyadic plane from its geometrical counterpart is equal to $\frac{1}{3} \log_2 n$, where n is linear size of the data cube.

ACKNOWLEDGEMENTS

Applied scientific research is supported by Ministry of Education and Science of the Russian Federation (projects RFMEFI58214X0002)

REFERENCES

- Ballester, P. 1994. "Hough transform for robust regression and automated detection." *Astronomy and Astrophysics*, 286:1011–1018.
- Bezmaternykh, P.V.; T.M. Khanipov; and D.P. Nikolaev. 2012. "Linear regression task solution with fast Hough transform (in Russian)." *35th Conference and School on Information Technologies and Systems (ITaS 2012)*, 354–359.
- Brady, M. 1998. "A fast discrete approximation algorithm for the Radon transform." *SIAM J. Computing*, 27(1):107–119.
- Ershov, E.I.; D.P. Nikolaev; V.V. Postnikov; and A.P. Terekhin. 2015a. "Exact fast algorithm for optimal linear separation of 2d distributions." *Proceedings 29th European Conference on Modelling and Simulation (ECMS 2015)*, 469–474.
- Ershov, E.I.; A.P. Terekhin; D.P. Nikolaev; V.V. Postnikov; and S.M. Karpenko, "Fast Hough transform analysis: pattern deviation from line segment." In *Eighth International Conference on Machine Vision*, 987509I 1–5 (International Society for Optics and Photonics, 2015b).
- Frederick, M.; N. VanderHorn; and A. Somani. 2005. "Real-time H/W implementation of the approximate discrete radon transform." *IEEE International Conference on Application-Specific Systems, Architecture Processors (ASAP'05)*, 2:399–404.
- Goldenshluger, A. and A. Zeevi. 2004. "The hough transform estimator." *The Annals of Statistics*, 32:19081932.
- Gotz, W.A. and H.J. Druckmiller. 1995. "A fast digital Radon transform – An efficient means for evaluating the Hough transform." *Pattern Recognition*, 28.12:1985–1992.
- Hart, P. 2009. "How the Hough transform was invented [DSP history]." *Signal Processing Magazine IEEE*, 26(6):18–22.
- Hart, P. and R. Duda. 1972. "Use of the Hough transformation to detect lines and curves in pictures." *Communications of the ACM*, 15:11–15.
- Hassanein, A.S.; S. Mohammad; M. Sameer; and M.E. Ragab. 2015. "A survey on Hough transform, theory, techniques and applications." *arXiv preprint arXiv:1502.02160*.
- Karpenko, S.M.; I.A. Konovalenko; A.B. Miller; B.M. Miller; and D.P. Nikolaev. 2015. "UAV Control on the Basis of 3D Landmark Bearing-Only Observations." *Sensors*, 15(12):29802–29820.
- Karpenko, S.M.; D.P. Nikolaev; P.P. Nikolayev; and V.V. Postnikov. 2004. "Fast Hough transform with controllable robustness (in Russian)." *In Proc. of IEEE AIS'04 and CAD-2004*, 2:303–309.
- Konovalenko, I.A.; A.B. Miller; B.M. Miller; and D.P. Nikolaev, "UAV navigation on the basis of the feature points detection on underlying surface." In *Proceedings of the 29th European Conference on Modeling and Simulation (ECMS 2015)*, Albena (Varna), Bulgaria, 499–505 (2015).
- Mukhopadhyay, P. and B.B. Chaudhuri. 2015. "A survey

of Hough transform." *Pattern Recognition*, 48(3):993–1010.

- Nikolaev, D.P.; S.M. Karpenko; I.P. Nikolaev; and P.P. Nikolayev. 2008. "Hough transform: underestimated tool in the computer vision field." *Proceedings of the 22th European Conference on Modelling and Simulation*, 238–246.

AUTHOR BIOGRAPHIES



EGOR ERSHOV was born in Moscow, USSR. He studied engineer science and mathematics, obtained his Master degree in 2014 from Moscow Institute of Physics and Technology. Now he is a Ph.D. student. Since 2014 he is working in Vision Systems Lab at the Institute for Information Transmission Problems RAS. His research activities are in the areas of computer vision. His e-mail address is ershov@iitp.ru.



VASILII POSTNIKOV was born in Sverdlovsk, USSR. He studied applied mathematics, obtained his Master degree in 1990 and Ph.D. degree in 2001 from Moscow Institute of Physics and Technology. Since 1993 he works at Institute for System Analysis, RAS. His research activities are in the area in the area of image analysis and video data recognition. His e-mail address is vasilli.postnikov@gmail.com.



SIMON KARPENKO was born in Moscow, USSR. He studied mathematics, obtained his Master degree in 2002 from Moscow State University. Since 2007 he works in the Vision Systems Lab. at the Institute for Information Transmission Problems RAS. His research activities are in the areas of computer vision, machine learning and pattern recognition. His e-mail address is simon@iitp.com.



ARSENIY TEREKHIN was born in Moscow, USSR. Since 2005 he worked as a programmer for a brokerage company. In 2014 he joined Vision Systems Lab. at the Institute for Information Transmission Problems, RAS. His e-mail address is ars@iitp.ru



DMITRY NIKOLAEV was born in Moscow, USSR. He studied physics and computer science, obtained his Master degree in 2000 and Ph.D. degree in 2004 from Moscow State University. Since 2007 he is a head of the Vision Systems Lab. at the Institute for Information Transmission Problems RAS. His research activities are in the areas of computer vision and image processing with focus on computation effective algorithms. His e-mail address is dimonstr@iitp.ru.

New Approach of Constant Resolving of Analytical Programming

Tomas Urbanek
Zdenka Prokopova
Radek Silhavy
Ales Kuncar

Department of Computer and Communication Systems
Tomas Bata University in Zlin
Nad Stranemi 4511
Email: turbanek@fai.utb.cz

KEYWORDS

Analytical Programming; Constant Creation; Differential Evolution; Symbolic Regression

ABSTRACT

This papers' aim is to provide the Artificial Intelligence community with a better tool for symbolic regression. In this paper, the method of analytical programming and constant resolving is revisited and extended. Nowadays, analytical programming mainly uses two methods for constant resolving. The first method is meta-evolution, in which the second evolutionary algorithm is used for constant resolving. The second method uses non-linear fitting algorithm. This paper reveals the third method, which use the basic mathematics to generate constants. The findings of this study have a number of important implications for future practice.

INTRODUCTION

A lot of application in symbolic regression field requires some kind of algorithm for constant generation (Koza 1992), (O'Neill, Brabazon & Ryan 2002). Therefore the effective numeric constant resolving algorithm is very important. In this paper, we use analytical programming as a method for symbolic regression (Zelinka, Davendra, Senkerik, Jasek & Oplatkova 2011). This method based on existence of any kind of evolutionary algorithm which generate a pointers to function tables. Analytical programming (AP) from these pointers maps and assemble a final regression function. In the analytical programming algorithm, two main approaches can be selected for constant resolving. The first approach is meta-evolution, in which the second or slave evolutionary algorithm is used for constant resolving. The second approach is to use of non-linear fitting algorithm. Both approaches added to analytical programming a lot of complexity. This study introduces a new approach for constant resolving in analytical programming algorithm. This approach is founded on basic mathematical calculation.

Section II is devoted to the original algorithm of analytical programming. Section III presents the new approach for constant resolving. Section IV presents the methods used for this study. Section V summaries the results of this research. Finally, Section VI presents the conclusions of this study.

Differential Evolution

Differential Evolution is an optimization algorithm introduced by Storn and Price in 1995, (Storn & Price 1995). This optimization method is an evolutionary algorithm based on population, mutation and recombination. Differential Evolution is easy to implement and has only four parameters which need to be set. The parameters are: Generations, NP, F and Cr. The Generations Parameter determines the number of generations; the NP Parameter is the population size; the F Parameter is the Weighting Factor; and the Cr Parameter is the Crossover Probability, (Storn 1996). In this research, the differential evolution is used as an analytical programming engine.

Analytical Programming

Analytical Programming, is a symbolic regression method. The core of analytical programming is a set of functions and operands. These mathematical objects are used for the synthesis of a new function. Every function in the analytical programming set core has its own varying number of parameters. The functions are sorted according to these parameters into General Function Sets (GFS). For example, GFS_{1par} contains functions that have only one parameter e.g. $\sin()$, $\cos()$, or other functions. AP must be used with any evolutionary algorithm that consists of a population of individuals for its run (Oplatkova, Senkerik, Zelinka & Pluhacek 2013).

The function of analytical programming can be seen in Figure 1. In this case, Evolutionary Algorithm is Differential Evolution. The initial population is generated using Differential Evolution. This population, which must consist of natural numbers, is used for analytical programming purposes. The analytical programming then constructs the function on the basis of this population. This function is evaluated by its Cost Function. If the termination condition is met, then the algorithm ends. If the condition is not met, then Differential Evolution creates a new population through the Mutation and Recombination processes. The whole process continues with the new population. At the end of the analytical programming process, it is assumed that one has a function that is the optimal solution for the given task.

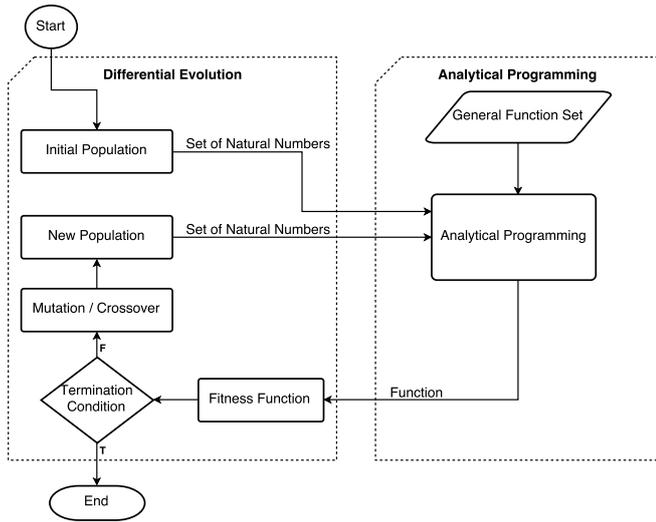


Fig. 1. Scheme of Analytical Programming with Differential Evolution algorithm

ORIGINAL ALGORITHM

Let's have the individual of the n length

$$\mathbf{ind} = (x_1, x_2, \dots, x_n)$$

where $x_i \in \mathbb{R}^+$.

This individual is then rounded

$$\mathbf{ind}_r = (\|x_1\|, \|x_2\|, \dots, \|x_n\|)$$

where $\|\cdot\|$ is nearest integer function.

Let's have a set called GFS_{all} which consists of m functions

$$GFS_{all} = \{\{f_1, fp_1\}, \{f_2, fp_2\}, \dots, \{f_m, fp_m\}\}$$

where f_m is function and fp_m is number of parameters of function f_m .

Then these functions are sorted to 4 sets : GFS_0 , $GFS_{1,0}$, $GFS_{2,1,0}$ and GFS_{all} .

TABLE I. GFS S BY PARAMETERS

$GFS_0 = \{\{f_1, 0\}, \{f_2, 0\}, \dots, \{f_n, 0\}\} \subset GFS_{all}$
$GFS_1 = \{\{f_1, 1\}, \{f_2, 1\}, \dots, \{f_n, 1\}\} \subset GFS_{all}$
$GFS_2 = \{\{f_1, 2\}, \{f_2, 2\}, \dots, \{f_n, 2\}\} \subset GFS_{all}$
$GFS_{1,0} = GFS_1 \cup GFS_0$
$GFS_{2,1,0} = GFS_2 \cup GFS_1 \cup GFS_0$

The sets from table I are expanded to the maximum value from the individual because we expected that the value of each number in individual could be higher then the length of GFS .

TABLE II. EXAMPLE OF GFS S, WHEN MAXIMUM VALUE OF INDIVIDUAL IS 6

$GFS_0 = \{\{K, 0\}, \{x, 0\}, \{K, 0\}, \{x, 0\}, \{K, 0\}, \{x, 0\}\}$
$GFS_{1,0} = \{\{Sin, 1\}, \{Cos, 1\}, \{K, 0\}, \{x, 0\}, \{Sin, 1\}, \{Cos, 1\}\}$
$GFS_{2,1,0} = \{\{Plus, 2\}, \{Minus, 2\}, \{Sin, 1\}, \{Cos, 1\}, \{K, 0\}, \{x, 0\}\}$
$GFS_{all} = \{\{Plus, 2\}, \{Minus, 2\}, \{Sin, 1\}, \{Cos, 1\}, \{K, 0\}, \{x, 0\}\}$

Then we need to construct a matrix with functions mapped by individual.

After that we have

$$\mathbf{function} = ((f_1, fp_1), (f_2, fp_2), \dots, (f_n, fp_n))$$

where f_n is function and fp_n is number of parameters of function f_n .

Then we need to choose which function is applied to which function. The values are pointers to the functions in GFS s. After that, we have constructed function; however, there is a constant K , which have to be resolved. Now we have two possibilities

- Meta-evolution e.g. Differential Evolution
- Non-linear least square fitting for example Levenberg-Marquardt

NEW APPROACH

Let's have the individual of the n length

$$\mathbf{ind} = (x_1, x_2, \dots, x_n)$$

where $x_i \in \mathbb{R}^+$.

This individual is then rounded

$$\mathbf{ind}_f = (\lfloor x_1 \rfloor, \lfloor x_2 \rfloor, \dots, \lfloor x_n \rfloor)$$

where $\lfloor \cdot \rfloor$ is round to floor.

Now we can construct a difference between \mathbf{ind} and \mathbf{ind}_f .

$$\mathbf{ind}_c = \mathbf{ind} - \mathbf{ind}_f$$

In vector \mathbf{ind}_f are pointers to GFS s and \mathbf{ind}_c are corresponding constants.

The decimal numbers in \mathbf{ind}_c are in range $< 0, 1 >$. These numbers could be easily converted to constants into the chosen range.

Let's have a set called GFS_{all} which consists of m functions

$$GFS_{all} = \{\{f_1, fp_1\}, \{f_2, fp_2\}, \dots, \{f_m, fp_m\}\}$$

where f_i is function and fp_i is number of parameters of function f_i .

Then this functions are sorted to 4 sets : GFS_0 , $GFS_{1,0}$, $GFS_{2,1,0}$ and GFS_{all} .

The next key change in analytical programming algorithm is function selection. In this new approach, the GFS s are not expanded to the maximum value of the individual. The selection of function is controlled by modulo function. On the position where the constant K will be mapped; we can read a constant number from the \mathbf{ind}_c vector. The original mapping algorithm is the same. Now we have constructed function with resolved constants.

PROBLEM STATEMENT

The overall research question to be answered within the study is whether there is a possibility to outperformed the original analytical programming method. This section presents the design of the research question. We performed experiments to get an insight in the constant resolving of analytical programming. The research question of our study can be outlined as follows:

RQ: Analysing the impact of new approach on the calculation duration and minimization performance of analytical programming.

The research question (RQ) aims to get an insight on the new approach of constant resolving of analytical programming and understand the actual effectiveness of this technique. For this reason, we use 3 different methods for constant resolving. Analytical programming with differential evolution and two new versions of proposed algorithm. Then, we try to outperformed the original constant resolving algorithm of analytical programming. To asses the performance of fitness function, we used descriptive statistics.

METHOD

New constant resolving algorithm for analytical programming was tested for searching regression functions. Results were compared by descriptive statistics.

Following functions have been tested

- $f(x) = 45.5$
- $f(x) = 3x + 0.65$
- $f(x) = 2.3x^2 - 20x - 5.6$
- $f(x) = 3.65 * \sin(2x)$

These functions were selected with emphasis on constant resolving. Functions such as constant, linear, quadratic and harmonic were tested. There was generated 20 points for each function. And the task for analytical programming was to fit these points. Three methods of constant resolving were tested.

- Analytical programming with differential evolution further referred to as AP+DE
- New analytical programming version with constant range $< -1000, 1000 >$ further referred to as AP2(-1000,1000)
- New analytical programming version with constant range $< 0, 10 >$ further referred to as AP2(0,10)

TABLE III. SYSTEM CONFIGURATION

Parameter	Value
CPU	AMD Phenom II X2 3GHz
RAM	8 GB
Operation system	Windows 7 Professional 64 bit
Programming language	LUA 5.2

Table III shows the system configuration for performing tests.

Table IV shows the analytical programming set-up. The number of leafs (functions built by analytical programming

TABLE IV. SET-UP OF ANALYTICAL PROGRAMMING

Parameter	Value
Number of leafs	16
GFS - functions	plus, minus, multiply, divide, power, log, log10, exp, sqrt, floor, ceil, abs, sin, cos
GFS - constants	x, K

can be seen as trees) was set to 16. This value was sufficient for the purpose of this paper.

TABLE V. SET-UP OF DIFFERENTIAL EVOLUTION

Parameter	Value
NP	45
Generations	300
F	0.2
Cr	0.9

Table V shows the set-up of differential evolution. The best set-up of differential evolution is the subject of further research.

A. Fitness function

Fitness function used for this task is as following:

$$LAD = \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

where y_i is actual value and \hat{y}_i is predicted value.

RESULTS

For each function, 100 equations were calculated by the aforementioned constant resolving approaches for statistical evaluation. There were collected information about duration and least absolute deviation error.

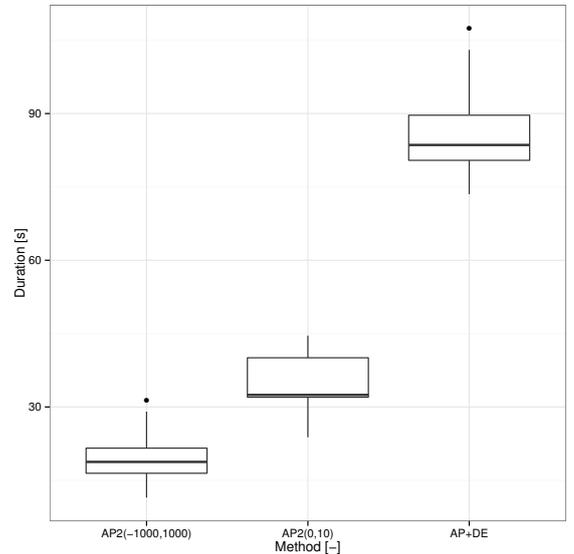


Fig. 2. Comparison of time duration of each method for function $f(x) = 45.5$

Figure 2 depicts the box plot comparison of time duration for each constant resolving method. As can be seen,

the new approach for constant resolving with constant range (-1000,1000) was nearly three times faster than analytical programming with differential evolution. AP2(0,10) performed slightly worse than AP2(-1000,1000).

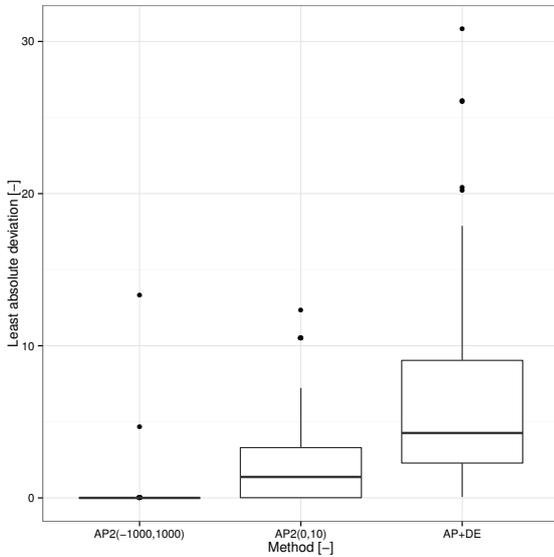


Fig. 3. Comparison of LAD error of each method for function $f(x) = 45.5$

Figure 3 depicts the box plot comparison of LAD error for each constant resolving method. As can be seen, AP2(-1000,1000) find the minimum nearly in each of 100 equations. All presented approaches find the minimum. AP2(0,10) method perform as the second best.

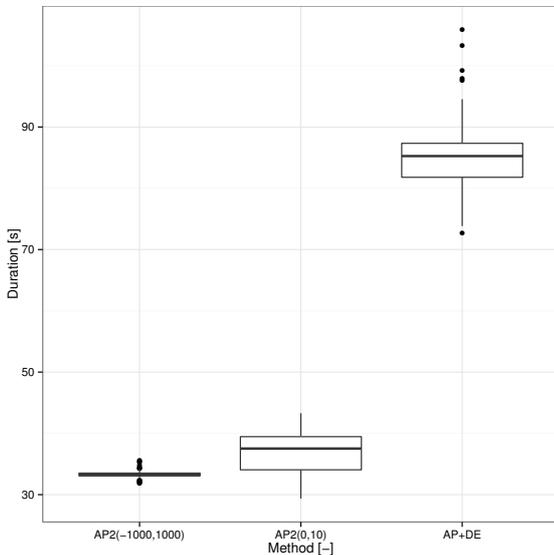


Fig. 4. Comparison of time duration of each method for function $f(x) = 3x + 0.65$

Figure 4 depicts the box plot comparison of time duration for each constant resolving method. As can be seen, AP2(-1000,1000) and AP2(0,10) performed nearly three times faster than analytical programming with differential evolution constant resolving. AP2(0,10) performed slightly worse than AP2(-1000,1000).

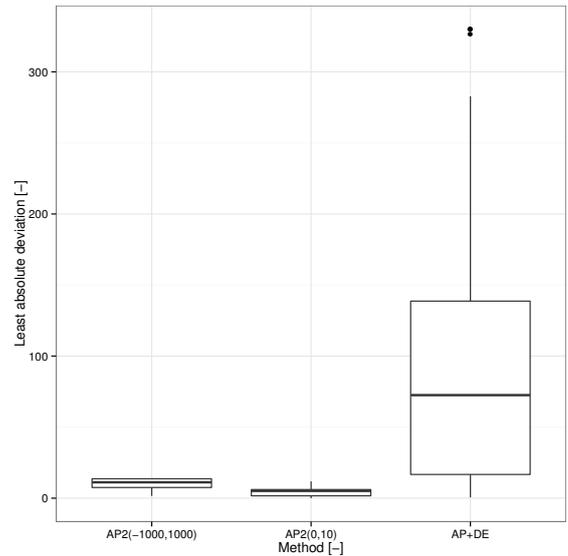


Fig. 5. Comparison of LAD error of each method for function $f(x) = 3x + 0.65$

Figure 5 depicts the box plot comparison of LAD error for each constant resolving method. As can be seen, AP2 generated lower error than 25% of AP+DE approach. There also can be seen a very low variance in AP2 method.

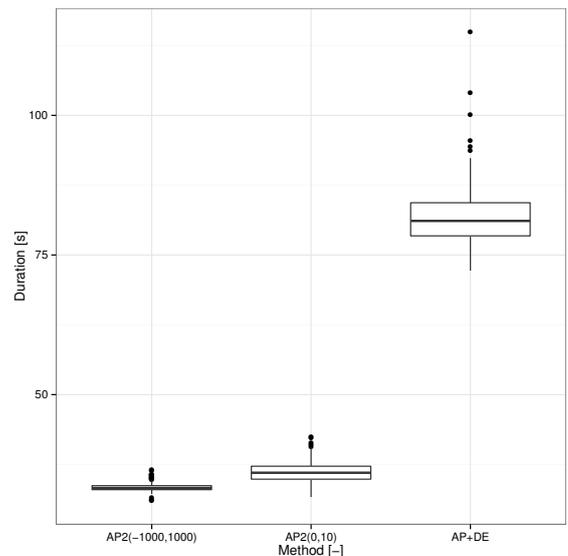


Fig. 6. Comparison of time duration of each method for function $f(x) = 2.3x^2 - 20x - 5.6$

Figure 6 depicts the box plot comparison of time duration for each constant resolving method. As can be seen, AP2(-1000,1000) and AP2(0,10) performed nearly two times faster than analytical programming with differential evolution constant resolving. AP2(-1000,1000) performed slightly worse than AP2(0,10).

Figure 7 depicts the box plot comparison of LAD error for each constant resolving method. None of the presented approaches find the minimum value. AP2(0,10) performed

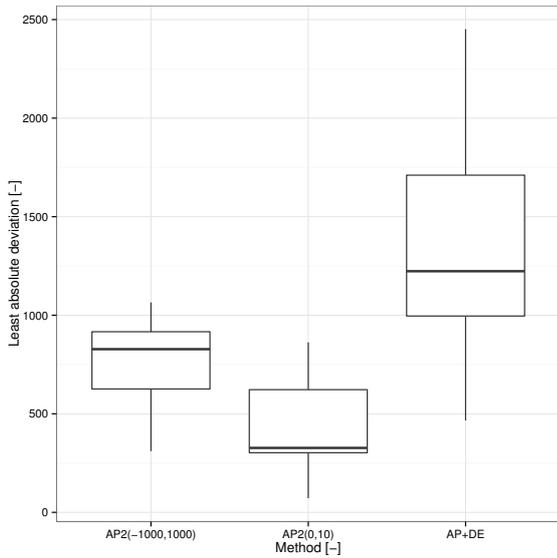


Fig. 7. Comparison of LAD error of each method for function $f(x) = 2.3x^2 - 20x - 5.6$

better than other presented approaches. As can be seen, AP2 approach had lower variance than AP+DE approach.

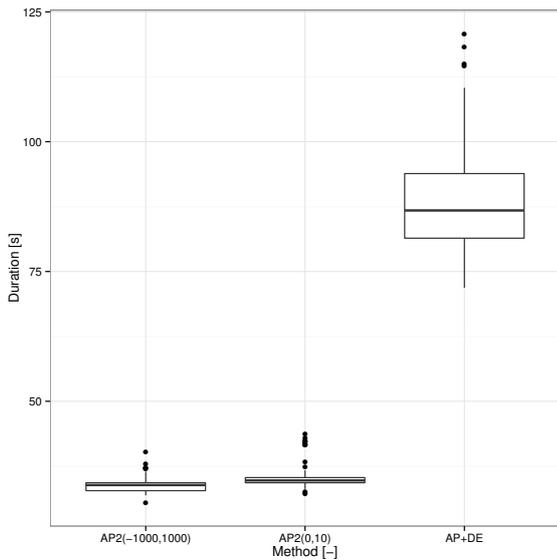


Fig. 8. Comparison of time duration of each method for function $f(x) = 3.65 * \sin(2x)$

Figure 8 depicts the box plot comparison of time duration for each constant resolving method. As can be seen, AP2(-1000,1000) and AP2(0,10) performed nearly three times faster than analytical programming with differential evolution constant resolving. AP2(-1000,1000) and AP2(0,10) performed nearly identical.

Figure 9 depicts the box plot comparison of LAD error for each constant resolving method. Only AP2(0,10) find the minimum value. AP+DE and AP2(-1000,1000) perform notably worse than AP2(0,10).

Table VI summarises the statistics for each method and

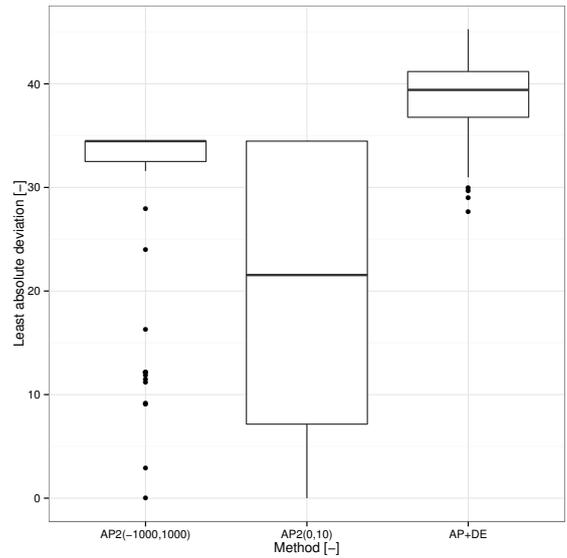


Fig. 9. Comparison of LAD error of each method for function $f(x) = 3.65 * \sin(2x)$

equation. As can be seen, the hardest equation to minimize was equation E3 where the minimum value was 71,84 for AP2(0,10). The new approach also has lower values for means and medians than meta-evolution approach AP+DE.

DISCUSSION

The study started out with a goal to answer the question of whether the new constant resolving technique outperforms the standard constant resolving solution in analytical programming algorithm. This question is answered in the result section.

There is question (RQ), which must be answered. For answering this question, we need to study figures in result section. As could be seen in figures 2, 4, 6 and 8, the new approach could achieve up to 3 times lower calculation duration than standard approaches. These results were expected, because we remove time complexity of constant resolving using another differential evolution. The most surprising aspect of the results is in the minimization performance. The figures 3, 5, 7 and 9 depicted that the minimization performance are more stable and the new approach finds more accurate minimum value; however, this could be caused by the setting of slave differential evolution.

THREATS OF VALIDITY

It is widely recognised that several factors can bias the validity of simulation studies. Therefore, our results are not devoid of validity threats.

External validity

External validity questions whether the results can be generalized outside the specifications of a study (Milicic & Wohlin 2004). The first validity issue to mention is that either analytical programming nor differential evolution has been exhausted via fine-tuning. Therefore, future work is required to exhaust all the parameters of these methods to

TABLE VI. STATISTICAL COMPARISON OF TESTED METHOD

Equation Method	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
E1 : AP2(-1000,1000)	0,00	0,00	0,00	0,18	0,00	13,36
E1 : AP2(0,10)	0,00	0,01	1,38	2,31	3,31	12,37
E1 : AP+DE	0,07	2,29	4,27	6,17	9,04	30,83
E2 : AP2(-1000,1000)	1,58	7,52	11,20	10,28	13,65	13,65
E2 : AP2(0,10)	0,00	1,66	5,12	4,38	6,07	11,87
E2 : AP+DE	0,62	16,70	72,52	93,85	138,68	330,06
E3 : AP2(-1000,1000)	310,80	625,90	827,90	798,00	916,30	1065,30
E3 : AP2(0,10)	71,84	302,40	326,83	433,71	622,45	862,60
E3 : AP+DE	466,10	996,20	1223,30	1345,80	1710,70	2451,50
E4 : AP2(-1000,1000)	0,01	32,50	34,47	31,17	34,47	34,47
E4 : AP2(0,10)	0,00	7,15	21,55	20,42	34,47	34,47
E4 : AP+DE	27,65	36,78	39,42	38,78	41,19	45,28

use their best versions. Threat to external validity could be also the implementation of the analytical programming and differential evolution algorithms. Although we used standard implementations, there is considerable amount of code, which could be the threat to validity.

CONCLUSIONS

In this paper, the new approach of constant resolving in analytical programming algorithm was presented. The presented approach is founded on basic mathematical calculations. The main benefit of this solution is that, there are no another calculation of evolutionary algorithm or non-linear fitting algorithm for constant resolving. There is also a benefit that there is no need to set a slave meta-evolution algorithm. Nevertheless, the range for constants must be set. Future research should therefore concentrate on the investigation of a proper set-up for analytical programming constant range.

ACKNOWLEDGEMENT

This study was supported by the internal grant of Tomas Bata University in Zlin No. IGA/FAI/2016/035 funded from the resources of specific university research.

REFERENCES

REFERENCES

- Koza, J. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press.
- Milicic, D. & Wohlin, C. (2004), Distribution patterns of effort estimations, *IEEE Conference Proceedings of Euromicro 2004, Track on Software Process and Product Improvement* pp. 422–429. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1333398
- O'Neill, M., Brabazon, A. & Ryan, C. (2002), *Genetic Algorithms and Genetic Programming in Computational Finance*, Springer US, Boston, MA, chapter Forecasting Market Indices Using Evolutionary Automatic Programming, pp. 175–195. http://dx.doi.org/10.1007/978-1-4615-0835-9_8
- Oplatkova, Z. K., Senkerik, R., Zelinka, I. & Pluhacek, M. (2013), Analytic programming in the task of evolutionary synthesis of a controller for high order oscillations stabilization of discrete chaotic systems, *Computers & Mathematics with Applications* 66(2), 177–189. <http://www.sciencedirect.com/science/article/pii/S0898122113001004>
- Storn, R. (1996), On the usage of differential evolution for function optimization, *Proceedings of North American Fuzzy Information Processing* pp. 519–523. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=534789>
- Storn, R. & Price, K. (1995), *Differential Evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces*, Vol. 11, Technical Report TR-95-012.

Zelinka, I., Davendra, D., Senkerik, R., Jasek, R. & Oplatkova, Z. (2011), *Analytical programming—a novel approach for evolutionary synthesis of symbolic structures*, InTech, Rijeka.

TOMAS URBANEK was born in Zlin in 1987. He received a B.Sc. (2009), M.Sc. (2011) in Information Technology from Faculty of Applied Informatics, Tomas Bata University in Zlin. He is a doctoral student at the Computer and Communication Systems Department. Major research interests are software engineering, effort estimation in software engineering and artificial intelligence

ZDENKA PROKOPOVA was born in Rimavska Sobota, Slovak Republic in 1965. She graduated from the Slovak Technical University in 1988, with a Masters degree in Automatic Control Theory. She received her Technical Cybernetics Doctoral degree in 1993 from the same university. She worked as an Assistant at the Slovak Technical University from 1988 to 1993. During 1993–1995, she worked as a programmer of database systems in the Datalock business firm. From 1995 to 2000, she worked as a Lecturer at Brno University of Technology. Since 2001, she has been at Tomas Bata University in Zlin, in the Faculty of Applied Informatics. She presently holds the position of Associate Professor at the Department of Computer and Communication Systems. Her research activities include programming and applications of database systems, mathematical modelling, computer simulation and the control of technological systems.

RADEK SILHAVY was born in Vsetin in 1980. He received a B.Sc. (2004), M.Sc. (2006), and Ph.D. (2009) in Engineering Informatics from the Faculty of Applied Informatics, Tomas Bata University in Zlin. He is a Senior Lecturer and researcher in the Computer and Communication Systems Department. His Ph.D. research was on The Verification of the Distributed Schema for the Electronic Voting System. His major research interests are software engineering, empirical software engineering and system engineering.

ALES KUNCAR was born in Prerov in 1989. He received a B.Sc. (2012), M.Sc. (2014) in Computer and Communication Systems from Faculty of Applied Informatics, Tomas Bata University in Zlin. He is a doctoral student at the Computer and Communication Systems Department. Major research interests are navigation systems, MEMS sensors and mathematical modelling.

ANALYTICAL PROGRAMMING WITH EXTENDED INDIVIDUALS

Adam Viktorin
Michal Pluhacek
Zuzana Kominkova Oplatkova
Roman Senkerik
Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{aviktorin, pluhacek, oplatkova, senkerik}@fai.utb.cz

KEYWORDS

Analytical Programming, Differential Evolution, SHADE, Constant Estimation

ABSTRACT

This paper proposes a new technique for the estimation of values of constants in programs synthesized by Analytical Programming (AP). Proposed technique divides the features of an individual in Evolutionary Algorithm (EA) into two parts – program part and constant part. Features in program part are mapped to synthesized program and features in constant part are used as constant values. AP with implementation of this technique is tested on four benchmark functions – Quintic, Sextic, 3Sine and 4Sine.

INTRODUCTION

Analytical Programming (AP) is a novel approach to symbolic structure synthesis which uses Evolutionary Algorithm (EA) for its computation. It was introduced by Zelinka in 2001 (Zelinka 2001) and since its introduction, it has been proven on numerous problems to be as suitable for symbolic structure synthesis as Genetic Programming (GP) (Koza 1990; Zelinka, Oplatkova 2003; Zelinka et al. 2005; Oplatkova, Zelinka 2006; Zelinka et al. 2008; Senkerik et al. 2013). AP is based on the set of functions and terminals called General Functional Set (GFS). The individual of an EA is translated from individual domain to program domain using this set. The subset of terminals often contains constants but the proper amount and which ones should be used is dependent on the optimized task. Zelinka et al. (Zelinka et al. 2005) solved this problem by using only one constant K , which is indexed in the synthesized formula as K_1, K_2, \dots, K_n and the values are estimated using secondary EA (meta-evolution) or by non-linear fitting. The first mentioned method is quite time consuming and the time estimation is hard because of the varying number of constants in synthesized programs. The later is also time consuming and depends on the implementation of the non-linear fitting method. Because of these disadvantages a new approach is presented in this paper which extends the individual, thus increasing the dimensionality. The extended features are used for the evolution of the constant values. Therefore, time complexity is increased, but it is

easy to estimate it and no secondary evolution process takes place. Furthermore, a novel state-of-art version of Differential Evolution (DE) called Success-History based Adaptive Differential Evolution (SHADE) (Tanabe, Fukunaga 2013) was used to carry out the evolution.

This approach was tested on some of the basic benchmark functions (Koza 1994) and the results are presented below.

The rest of the paper is structured as follows: Next section provides the description of AP, section that follows describes the constant handling process and following section is about DE and SHADE. Experiment setup and results are provided in two following sections and the paper is concluded in the last section.

ANALYTICAL PROGRAMMING

The basic functionality of AP is formed by three parts – General Functional Set (GFS), Discrete Set Handling (DSH) and Security Procedures (SPs). GFS contains all elementary objects which can be used to form a program, DSH carries out the mapping of individuals to programs and SPs are implemented into mapping process to avoid mapping to pathological programs and into cost function to avoid critical situations.

General Functional Set

AP uses sets of functions and terminals. Functions require at least one additional argument for computation, whereas terminals require no arguments and are final (e.g. constants, independent variables). The synthesized program is branched by functions requiring two and more arguments and the length of it is extended by functions which require one argument. Terminals do not contribute to the complexity of the synthesized program (length) but are needed in order to synthesize a non-pathological program (program that can be evaluated by cost function). Therefore, each non-pathological program must contain at least one terminal. Combined set of functions and terminals forms GFS which is used for mapping from individual domain to program domain. The content of GFS is dependent on user choice. GFS is nested and can be divided into subsets according to the number of arguments that the subset requires. GFS_{0arg} is a subset which requires zero arguments, thus contains only terminals. GFS_{1arg} contains all terminals and functions requiring one

argument, GFS_{2arg} contains all objects from GFS_{1arg} and functions requiring two arguments and so on, GFS_{all} is a complete set of all elementary objects. The GFS used in the theoretical part of this paper is depicted below and the division into subsets is also shown.

- GFS_{0arg} : x, k
- GFS_{1arg} : \sin, \cos, x, k
- $GFS_{2arg} = GFS_{all}$: $+, -, *, /, \sin, \cos, x, k$

For the purpose of mapping from individual to the program, it is important to note that objects in GFS are ordered by a number of arguments they require in descending order.

Discrete Set Handling

DSH is used for mapping the individual to the synthesized program. Most of the EAs use individuals with real numbered features. The first important step in order for DSH to work is to get individual with integer features which is done by rounding real feature values. The integer values of an individual are indexes into the discrete set, in this case, GFS and its subsets. If the index value is greater than the size of used GFS, modulo operation with the size of the discrete set is performed. Two examples of mapping are depicted in (1) and (2).

$$\begin{aligned} Individual &= \{0.12, 4.29, 6.92, 6.12, 2.45, \\ &\quad \{6.33, 5.78, 0.22, 1.94, 7.32\} \\ Rounded\ individual &= \{0, 4, 7, 6, 2, 6, 6, 0, 2, 7\} \quad (1) \\ GFS_{all} &= \{+, -, *, /, \sin, \cos, x, k\} \\ Program &: \sin x + k \end{aligned}$$

The objects in GFS_{all} are indexed from 0 and mapping is as follows: The first rounded individual feature is 0 which represents $+$ function in GFS_{all} . This function requires two arguments and those are represented by next two indexes $- 4$ and 7 , which are mapped to function \sin and constant k . The \sin function requires one argument which is given by next index (rounded feature) $- 6$ and it is mapped to variable x . Since there is no possible way of branching the program further, other features are ignored and synthesized program is $\sin x + k$.

Mapping steps:

1. Index 0 is mapped to $+$. Program: $_ + _$
2. Index 4 is mapped to \sin . Program: $\sin _ + _$
3. Index 7 is mapped to k . Program: $\sin _ + k$
4. Index 6 is mapped to x . Program: $\sin x + k$
5. Remaining indexes $\{2, 6, 6, 0, 2, 7\}$ are ignored because the program is complete.

Where $_$ denotes the space in the program which needs to be filled with objects from GFS.

$$\begin{aligned} Individual &= \{5.08, 1.64, 5.58, 4.41, 6.20, \\ &\quad \{1.28, 0.07, 3.99, 5.27, 2.64\} \\ Rounded\ individual &= \{5, 2, 6, 4, 6, 1, 0, 4, 5, 3\} \quad (2) \\ GFS_{all} &= \{+, -, *, /, \sin, \cos, x, k\} \\ Program &: \cos(x * \sin x) \end{aligned}$$

The first index to GFS_{all} is 5, which represents \cos function, its argument is chosen by next index $- 2$ representing function $*$ which needs two arguments.

Arguments are indexed 6 and 4 – variable x and function \sin . In this step only one more argument for function \sin is needed and it is variable x denoted by index 6. The synthesized program is therefore $\cos(x * \sin x)$.

Mapping steps:

1. Index 5 is mapped to \cos . Program: $\cos _$
2. Index 2 is mapped to $*$. Program: $\cos(_ * _)$
3. Index 6 is mapped to x . Program: $\cos(x * _)$
4. Index 4 is mapped to \sin . Program: $\cos(x * \sin _)$
5. Index 6 is mapped to x . Program: $\cos(x * \sin x)$
6. Remaining indexes $\{1, 0, 4, 5, 3\}$ are ignored because the program is complete.

It is worthwhile to note that in both examples individual features were not fully used and synthesized programs are not as complex as the dimensionality enables them to be. Moreover, both examples use indexes which are lower than the size of GFS_{all} therefore, no modulo operation is needed.

Security Procedures

SPs are used in AP to avoid critical situations. Some of the SPs are implemented into the AP itself and some have to be implemented into the cost function evaluation. The typical representatives of the later are checking synthesized programs for loops, infinity and imaginary numbers if not expected (dividing by 0, square root of negative numbers, etc.).

The most significant SP implemented in AP is checking for pathological programs. Pathological programs are programs which cannot be evaluated due to the absence of arguments in the synthesized function. For example, individual with rounded features of $\{4, 4, 4, 4, 4\}$ would be mapped to program $\cos(\cos(\cos(\cos(\cos _))))$ which lacks constant or variable at the empty position denoted by $_$ and thus represent a pathological program. Such situation can be avoided by a simple procedure which checks how far is the end of the individual during mapping and according to that maps rounded individual features to subsets of GFS_{all} which do not require too many arguments. With the previous example using GFS from GFS section, the mapping process would be as follows:

1. First three features $\{4, 4, 4\}$ already mapped to GFS_{all} . Program: $\cos(\cos(\cos _))$
2. Current index in rounded individual features is 4 and only 1 feature is left to the end of the individual therefore, the index is mapped to GFS_{1arg} and modulo operation by the size of GFS_{1arg} which is 4 is performed, thus $index = 4 \bmod 4 = 0$. The index is mapped to \sin . Program: $\cos(\cos(\cos(\sin _)))$
3. Last index in rounded individual is 4 and no features are left to the end of the individual, therefore index is mapped to GFS_{0arg} and modulo operation by the size of GFS_{0arg} which is 2 is performed, thus $index = 4 \bmod 2 = 0$. Index is mapped to x . Program: $\cos(\cos(\cos(\sin x)))$

The program is no longer pathological and can be evaluated. This simple SP is able to eliminate the generation of pathological programs and, therefore, improve the performance of AP because all individuals can be evaluated.

CONSTANT HANDLING

The main novelty of this paper is in the constant handling technique used with AP. In the previous work (Zelinka et al. 2005), constant values in synthesized programs were estimated by second EA (meta-heuristic) or by non-linear fitting. This work presents a novel approach, which uses the extended part of the individual in EA for the evolution of constant values.

The important task was to determine, what is the correct size of an extension (3).

$$k = l - \text{floor}((l - 1) / \text{max_arg}) \quad (3)$$

Where k is the maximum number of constants that can appear in the synthesized program (extension) of length l and max_arg is the maximum number of arguments needed by functions in GFS. Also the $\text{floor}()$ is a common floor round function. The final individual dimensionality (length) will be $k+l$ and the example might be:

- Program length $l = 10$
- GFS: $\{+, -, *, /, \sin, \cos, x, k\}$
- GFS maximum argument $\text{max_arg} = 2$
- Extension size $k = 10 - \text{floor}((10-1) / 2) = 6$
- Dimensionality of the extended individual $k+l = 16$

This means, that the EA will work with individuals of length 16, but only first 10 features will be used for indexing into the GFS and the rest will be used as constant values.

While mapping the individual into a program, the constants are indexed and later replaced by the value from individual. Simple example can be seen in (4). Individual features in bold are the constant values. It is worthwhile to note that only features which are going to be mapped to GFS are rounded and the rest is omitted.

$$\begin{aligned} \text{Individual} &= \{ 5.08, 1.64, 6.72, 1.09, 6.20, \} \\ &\{ 1.28, \mathbf{0.07}, \mathbf{3.99}, \mathbf{5.27}, \mathbf{2.64} \} \\ \text{Rounded individual} &= \{5, 2, 7, 1, 6, 1\} \\ \text{GFS}_{\text{all}} &= \{+, -, *, /, \sin, \cos, x, k\} \\ \text{Program: } &\cos(k1 * (x - k2)) \\ \text{Replaced: } &\cos(\mathbf{0.07} * (x - \mathbf{3.99})) \end{aligned} \quad (4)$$

The first index to GFS_{all} is 5, which represents \cos function, its argument is chosen by next index – 2 representing function $*$ which needs two arguments. Arguments are indexed 7 and 1 – constant $k1$ and function $-$. After this step, two arguments are needed and only two features are left in the program part of the individual. Therefore, the security procedure takes place and those last two features are indexed into GFS_{0arg} . Thus indexes 6 and 1 are mapped to variable x (6 mod

$\text{size}(\text{GFS}_{\text{0arg}}) = 0$) and constant $k2$. The synthesized program is therefore $\cos(k1 * (x - k2))$. The constants are replaced by the remaining features 0.07 and 3.99 respectively.

Individual mapping steps:

1. Index 5 is mapped to \cos . Program: $\cos(_ * _)$
2. Index 2 is mapped to $*$. Program: $\cos(_ * _)$
3. Index 7 is mapped to $k1$. Program: $\cos(k1 * _)$
4. Index 1 is mapped to $-$. Program: $\cos(k1 * (_ - _))$
5. Index 6 mod 2 = 0 is mapped to x . Program: $\cos(k1 * (x - _))$
6. Index 1 is mapped to $k2$. Program: $\cos(k1 * (x - k2))$
7. Constants are replaced by the remaining features according to their indexes – $k1$ is replaced by the first available feature 0.07 and $k2$ is replaced by the second available feature 3.99. Program: $\cos(0.07 * (x - 3.99))$

SUCCESS-HISTORY BASED ADAPTIVE DIFFERENTIAL EVOLUTION

This section describes the basics of DE and SHADE algorithms.

DE algorithm has three control parameters – population size NP , crossover rate CR and scaling factor F . In the canonical form of DE, those three parameters are static and depend on the user setting. Other important features of DE algorithm are mutation strategy and crossover strategy. Canonical DE uses “rand/1/bin” mutation strategy (5) and binomial crossover (8). SHADE algorithm, on the other hand, uses only two control parameters – population size NP and size of historical memories H . F and CR parameters are automatically adapted based on the evolutionary process and its values for each individual are generated according to (7) and (9) respectively. Also, the mutation strategy is different than that of canonical DE. Novel mutation strategy used in SHADE is called “current-to- p best/1” and it is depicted in (6). The concept of basic operations in DE and SHADE algorithms is shown in following sections, for a detailed description on feature constraint correction, update of historical memories and external archive handling in SHADE see (Tanabe, Fukunaga 2013).

Mutation Strategies and Parent Selection

In canonical forms of both algorithms, parent vectors are selected by classic PRNG with uniform distribution. Mutation strategy “rand/1/bin” uses three random parent vectors with indexes $r1$, $r2$ and $r3$, where $r1 = U[1, NP]$, $r2 = U[1, NP]$, $r3 = U[1, NP]$ and $r1 \neq r2 \neq r3$. Mutated vector $\mathbf{v}_{i,G}$ is obtained from three different vectors \mathbf{x}_{r1} , \mathbf{x}_{r2} , \mathbf{x}_{r3} from current generation G with help of static scaling factor $F_i = F$ as follows:

$$\mathbf{v}_{i,G} = \mathbf{x}_{r1,G} + F_i(\mathbf{x}_{r2,G} - \mathbf{x}_{r3,G}) \quad (5)$$

Contrarily, SHADEs mutation strategy “current-to- p best/1” uses four parent vectors – current i -th vector

$\mathbf{x}_{i,G}$, vector $\mathbf{x}_{pbest,G}$ randomly selected from $NP \times p$ ($p = U[p_{min}, 0.2]$, $p_{min} = 2/NP$) best vectors (in terms of objective function value) from G , randomly selected vector $\mathbf{x}_{r1,G}$ from G and randomly selected vector $\mathbf{x}_{r2,G}$ from the union of G and external archive A . Where $\mathbf{x}_{i,G} \neq \mathbf{x}_{r1,G} \neq \mathbf{x}_{r2,G}$.

$$\mathbf{v}_{i,G} = \mathbf{x}_{i,G} + F_i(\mathbf{x}_{pbest,G} - \mathbf{x}_{i,G}) + F_i(\mathbf{x}_{r1,G} - \mathbf{x}_{r2,G}) \quad (6)$$

The scaling factor F_i is generated from Cauchy distribution with location parameter value of $M_{F,r}$ which is randomly selected value from scale factor historical memory, and scale parameter value of 0.1 (7).

$$F_i = C[M_{F,r}, 0.1] \quad (7)$$

Crossover and Elitism

The trial vector $\mathbf{u}_{i,G}$ which is compared with original vector $\mathbf{x}_{i,G}$ is completed by crossover operation (8) and this operation is the same for both DE and SHADE algorithms. CR_i value in DE algorithm is again static $CR_i = CR$ whereas with SHADE algorithm its value is generated from a normal distribution with a mean parameter value of $M_{CR,r}$ which is randomly selected value from crossover rate historical memory and with standard deviation value of 0.1 (9).

$$\mathbf{u}_{j,i,G} = \begin{cases} \mathbf{v}_{j,i,G} & \text{if } U[0,1] \leq CR_i \text{ or } j = j_{rand} \\ \mathbf{x}_{j,i,G} & \text{otherwise} \end{cases} \quad (8)$$

Where j_{rand} is randomly selected index of a feature, which has to be updated ($j_{rand} = U[1, D]$), D is the dimensionality of the problem.

$$CR_i = N[M_{CR,r}, 0.1] \quad (9)$$

Vector which will be in the next generation $G+1$ is selected by elitism. When the objective function value of trial vector $\mathbf{u}_{i,G}$ is better than that of the original vector $\mathbf{x}_{i,G}$, the trial vector will be selected for the next population and the original will be placed into the external archive A . Otherwise, the original will survive and the content of A remains unchanged (10).

$$\mathbf{x}_{i,G+1} = \begin{cases} \mathbf{u}_{i,G} & \text{if } f(\mathbf{u}_{i,G}) < f(\mathbf{x}_{i,G}) \\ \mathbf{x}_{i,G} & \text{otherwise} \end{cases} \quad (10)$$

EXPERIMENT SETTING

The regression capabilities of proposed version of the AP algorithm were tested on four typical benchmark functions used in (Koza 1994):

- Quintic: $y = x^5 - 2x^3 + x$
- Sextic: $y = x^6 - 2x^4 + x^3$
- 3Sine: $y = \sin x + \sin 2x + \sin 3x$
- 4Sine: $y = \sin x + \sin 2x + \sin 3x + \sin 4x$

The dataset of 50 points for Quintic and Sextic functions was generated in x range $\langle -1, 1 \rangle$. Also 50 points from x range $\langle -\pi, \pi \rangle$ were generated for function 3Sine and 4Sine. All four datasets are displayed in Figures 1, 2, 3 and 4.

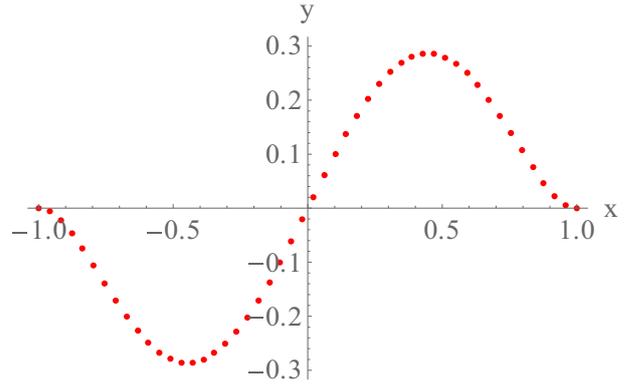


Figure 1: Quintic Dataset

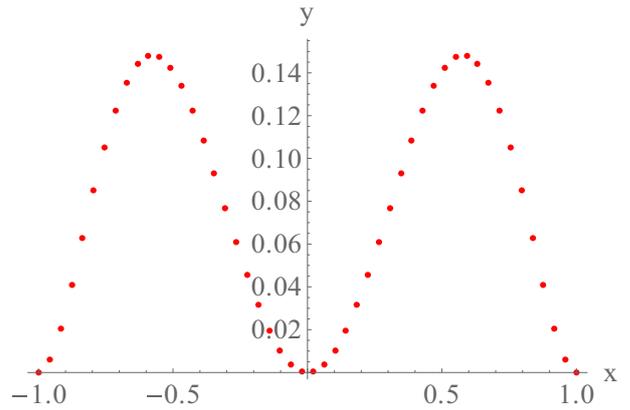


Figure 2: Sextic Dataset

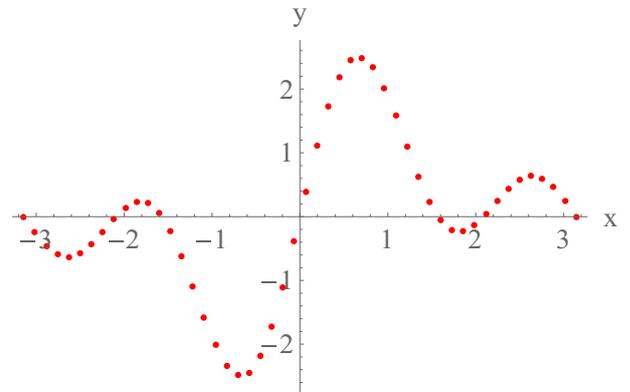


Figure 3: 3Sine Dataset

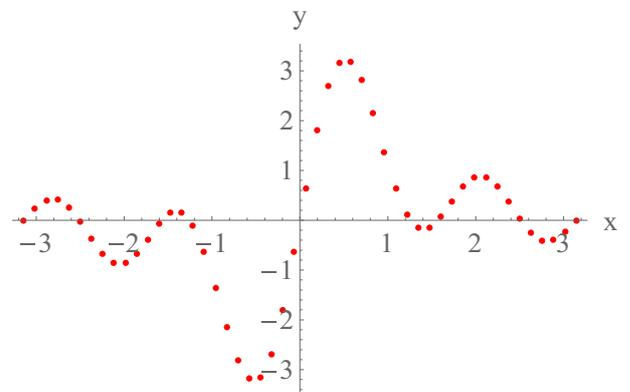


Figure 4: 4Sine Dataset

The settings of AP for the experiments and the settings of the underlying EA – SHADE can be found in Table 1.

Table 1: AP and SHADE Settings

	Quintic, Sextic	3Sine, 4Sine
AP		
GFS	+, -, *, /, x, k	
SHADE		
Dimension	90 (60 program + 30 constants)	180 (120 program + 60 constants)
Population size	50	100
Generations	4000	2000
Historical memory size	50	100

All datasets were synthesized 30 times and the results are presented in the next section.

RESULTS

Simple descriptive statistic results over 30 independent runs on all four datasets are given in Table 2. The table shows minimum, maximum, mean, median and standard deviation values obtained throughout the test. The fitness value of a synthesized program is a sum of the absolute distances between dataset points and their synthesized equivalents. Best results are displayed in Figures 5, 6, 7 and 8. The dataset is illustrated by red points and the best solution is represented by a green line.

Table 2: Results Over 30 Independent Runs of AP with SHADE and Extended Individuals

Dataset	Min	Max	Mean	Med	StD
Quintic	0.21	6.36	1.60	0.61	2.19
Sextic	0.24	2.37	0.75	0.62	0.54
3Sine	11.96	37.47	14.94	12.79	6.18
4Sine	14.23	26.09	16.54	15.77	2.57

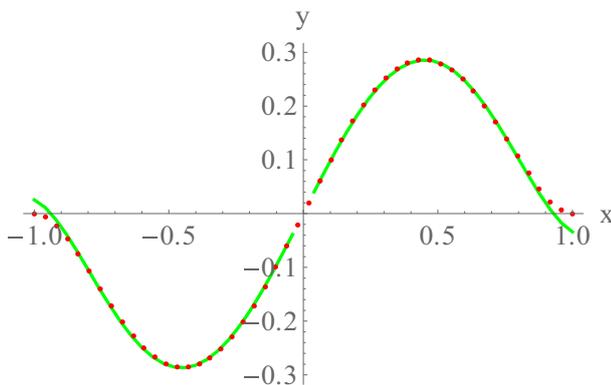


Figure 5: Quintic Regression

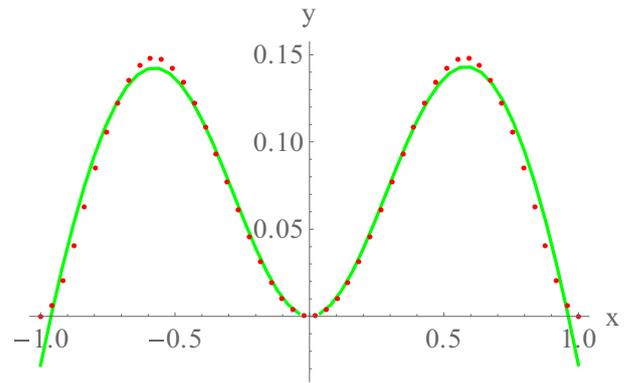


Figure 6: Sextic Regression

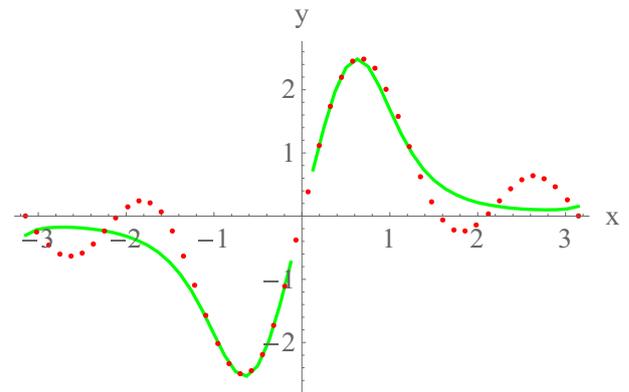


Figure 7: 3Sine Regression

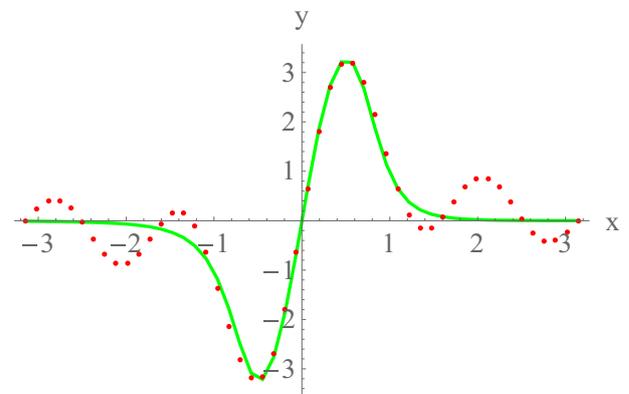


Figure 8: 4Sine Regression

As expected from the results in Table 2, the synthesized functions for 3Sine and 4Sine dataset are far less precise in modeling, than that of Quintic and Sextic datasets. In order to obtain a better model for 3Sine and 4Sine datasets, the content of GFS was extended by *sin* function and the most precise models are shown in Figure 9 and Figure 10.

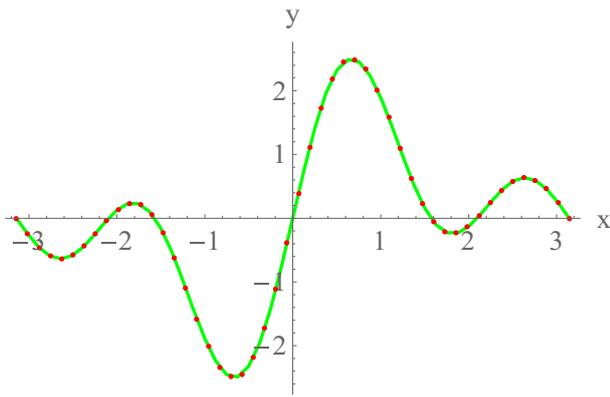


Figure 9: 3Sine Regression with *sin* in GFS

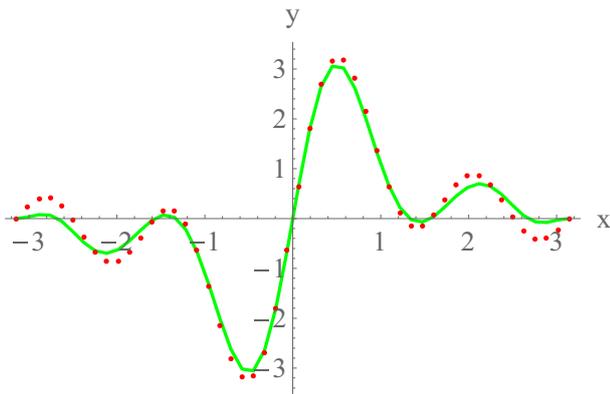


Figure 10: 4Sine Regression with *sin* in GFS

CONCLUSION

This paper presented a novel approach to estimation of the constant values in programs synthesized by AP. This approach uses part of the features of an individual in EA for the evolution of constant values.

As can be seen from the results, the performance of this variant is sufficient for two of the four benchmark test functions – Quintic and Sextic. Those two datasets are generated by polynomials and therefore can be synthesized from the functions and terminals contained in the GFS. Contrarily, 3Sine and 4Sine datasets are generated by harmonic functions which are not present in the GFS. This leads to the less precise regression, but the main trend is still transcribed. Further experiments confirmed, that when the harmonic function *sin* is added into the GFS, AP is able to find a better solution. This opens an interesting topic for further research in the possibility of using AP variant with extended individuals for the prediction of a suitable content of GFS. Determining the most suitable content of GFS for real world problems is a complex task and therefore a technique for prediction might be beneficial.

All four datasets tested in this study were taken from (Koza 1994). The proposed variant uses evolution for estimating the constant values, creating an infinite space of possibly synthesized programs and thus making it less suitable for non-complex test functions where the static value constants would be more reasonable

(Quintic, Sextic, 3Sine, 4Sine, e. g.). The comparison between GP and AP on those datasets can be found in (Zelinka, Oplatkova 2003). One of the goals of this paper was to demonstrate that AP with extended individuals is still able to perform a regression which transcribes the original trend of a non-complex dataset, but the main area of use is for the regression of complex datasets where the continuous constant space is necessary.

The advantage of the proposed approach is in its time complexity which is increased against canonical AP only by the increase in size of an individual in used EA. While other approaches (meta-evolution and non-linear fitting) might provide more precise results, their time complexity is higher and a lot harder to estimate which might be a problem in real time regression and prediction. The time complexity of meta-evolution is dependent on settings of the secondary EA and the time complexity of the non-linear fitting approach strongly depends on the used fitting algorithm and on the number of constants to estimate.

Further research will be targeted at the use of proposed variant of AP in real world applications where the constant estimation will be more likely to bring benefits in regression and at improving its precision.

ACKNOWLEDGEMENT

This work was supported by Grant Agency of the Czech Republic – GACR P103/15/06700S, further by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014). Also by the 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 Projects no. IGA/CebiaTech/2016/007.

REFERENCES

- Koza, J. R. 1990. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford University, Department of Computer Science.
- Koza, J. R. 1994. Genetic programming II: Automatic discovery of reusable subprograms. Cambridge, MA, USA.
- Oplatková, Z. and Zelinka, I. 2006. Investigation on artificial ant using analytic programming. In Proceedings of the 8th annual conference on Genetic and evolutionary computation (pp. 949-950). ACM.
- Senkerik, R.; Oplatková, Z.; Zelinka, I. and Davendra, D. 2013. Synthesis of feedback controller for three selected chaotic systems by means of evolutionary techniques: Analytic programming. Mathematical and Computer Modelling, 57(1), 57-67.
- Tanabe, R. and Fukunaga, A. 2013. Success-history based parameter adaptation for differential evolution. In Evolutionary Computation (CEC), 2013 IEEE Congress on (pp. 71-78). IEEE.
- Zelinka, I. 2001. Analytic programming by means of new evolutionary algorithms, Proceedings of 1st International Conference on New Trends in Physics'01, Brno, Czech Republic, pp. 210-214.
- Zelinka, I. and Oplatkova, Z. 2003. Analytic programming – comparative study, Proceedings of Second International

Conference on Computational Intelligence, Robotics, and Autonomous Systems, Singapore.

Zelinka, I.; Chen, G. and Celikovsky, S. 2008. Chaos synthesis by means of evolutionary algorithms. *International Journal of Bifurcation and Chaos*, 18(04), 911-942.

Zelinka, I.; Oplatkova, Z. and Nolle, L. 2005. Analytic programming–Symbolic regression by means of arbitrary evolutionary algorithms. *Int. J. of Simulation, Systems, Science and Technology*, 6(9), 44-56.

AUTHOR BIOGRAPHIES

ADAM VIKTORIN was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the field of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@fai.utb.cz



MICHAL PLUHACEK was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011 and Ph.D. in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. He now works as a researcher at the same university. His email address is: pluhacek@fai.utb.cz



ZUZANA KOMINKOVA OPLATKOVA is an associate professor at Tomas Bata University in Zlín. 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 email address is: oplatkova@fai.utb.cz



ROMAN SENKERIK was born in the Czech Republic, and went to the Tomas Bata University in Zlín, 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



MULTI-CHAOTIC DIFFERENTIAL EVOLUTION FOR VEHICLE ROUTING PROBLEM WITH PROFITS

Adam Viktorin
Dusan Hrabec
Michal Pluhacek

Tomas Bata University in Zlin, Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{aviktorin, hrabec, pluhacek}@fai.utb.cz

KEYWORDS

Vehicle Routing Problem with Profits, Differential Evolution, Optimization, Chaos

ABSTRACT

In this paper a new multi-chaotic variant of differential evolution is used to solve a model of vehicle routing problem with profits. The main goal was to achieve exceptional reliability (success rate) and low time demands in comparison with deterministic solvers. The method will be applied in the future on solving real-world transportation network problems.

INTRODUCTION

Various vehicle routing problem (VRP) variants are still actual category of optimization problems in these days and their solving is challenging for many optimization methods (Laporte 1992; Boussier et al 2007; Avci, Topaloglu 2016). However, in this paper we deal with a modification of VRP (or travelling salesman problem (TSP) alternatively) that is the so-called VRP with profits (the VRP with profits on a non-complete graph or selective VRP, alternatively), where not all customers have to be visited, see (Boussier et al, 2007).

We present initial results of evolutionary optimization method called multi-chaotic success-history based adaptive differential evolution that is being developed for future application on real vehicle routing problems (Pavlas et.al, 2015; Stodola et.al., 2014) and transportation network problems (Roupec et. al., 2013). Approach that we present in the paper is considered to be further developed to follow ideas in network design problems (Roupec et. al., 2013) and in waste management (Somplak et. al., 2013). Other modifications of the problem as well as algorithm are also considered (Stodola et.al., 2014).

The differential evolution (DE) (Storn, Price, 1997) is a foundation for some of the best performing evolutionary optimizers. In recent years, various successful applications of DE enhanced with chaotic pseudo-random number generators (PRNGs) were presented (Senkerik et.al., 2013, Liang et.al., 2011).

In the following section the basics of chaotic systems (maps) and their use as PRNGs are presented. The next section describes proposed Multi-chaotic differential evolution algorithm. In the following section the

problem is defined following with the experiment setup. The results are presented and discussed in the following section. The paper closes with a conclusion section.

CHAOTIC MAPS

Chaotic maps are systems generated continuously by simple equations from a single initial position. The current position is used for generation of a new position thus creating a sequence which is extremely sensitive to the initial position, which is also known as the “butterfly effect.” Sequences generated by chaotic maps have characteristics which are not common in classical random number generation. Therefore, their application in evolutionary algorithm (EA) can change its behavior and improve the performance.

The multi-chaotic system presented in this paper uses five different chaotic maps – Burgers, Delayed Logistic, Dissipative, Lozi and Tinkerbell.

The process of acquiring i -th random integer $rndInt_i$ from chaotic map is depicted in (1).

$$rndInt_i = \text{round}\left(\frac{\text{abs}(X_i)}{\max(\text{abs}(X_{i \in N}))} * (\text{maxRndInt} - 1)\right) + 1 \quad (1)$$

Where $\text{abs}(X_i)$ is an absolute value of i -th X of a chaotic sequence with length of N , $\max(\text{abs}(X_{i \in N}))$ is a maximum of absolute values of X in chaotic sequence and $\text{round}()$ is common rounding function. The generated number $rndInt_i$ is from interval $[1, \text{maxRndInt}]$.

Lozi Chaotic Map

The Lozi map is a simple discrete two-dimensional chaotic map. The map equations are given in (2). The typical parameter values are: $a = 1.7$ and $b = 0.5$ with respect to (Spratt, 2013). 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. The x,y plot of Lozi map with the typical setting is depicted in Figure 1.

$$\begin{aligned} X_{n+1} &= 1 - a|X_n| + bY_n \\ Y_{n+1} &= X_n \end{aligned} \quad (2)$$

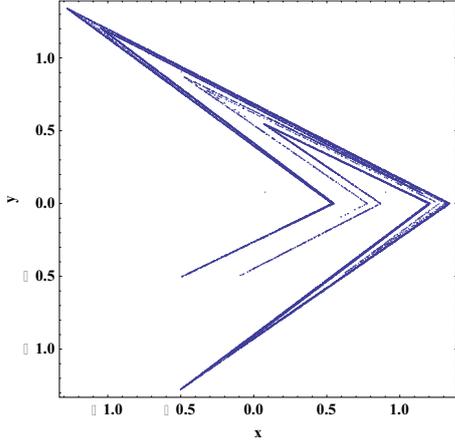


Figure 1: x,y plot of Lozi map

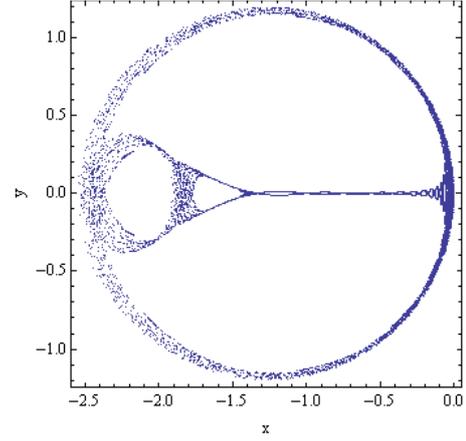


Figure 3: x,y plot of Burgers map

Dissipative Standard Map

The Dissipative standard map is a two-dimensional chaotic map. The parameters used in this work are $b = 0.6$ and $k = 8.8$ based on previous experiments [15, 16] and suggestions in literature (Spratt, 2013). The x,y plot of Dissipative standard map is given in Figure 2. The map equations are given in (3).

$$\begin{aligned} X_{n+1} &= X_n + Y_{n+1} \pmod{2\pi} \\ Y_{n+1} &= bY_n + k \sin X_n \pmod{2\pi} \end{aligned} \quad (3)$$

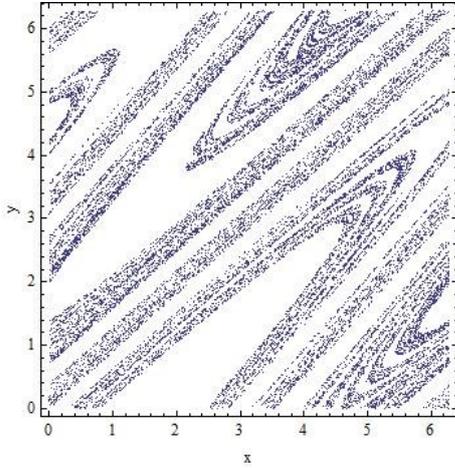


Figure 2: x,y plot of Dissipative standard map

Burgers Chaotic Map

The Burgers map (See Fig. 3) is a discretization of a pair of coupled differential equations. The map equations are given in (4) with control parameters $a = 0.75$ and $b = 1.75$ as suggested in (Spratt, 2013).

$$\begin{aligned} X_{n+1} &= aX_n - Y_n^2 \\ Y_{n+1} &= bY_n + X_n Y_n \end{aligned} \quad (4)$$

Tinkerbell Map

The Tinkerbell map is a two-dimensional complex discrete-time dynamical system given by (5) with following control parameters: $a = 0.9$, $b = -0.6$, $c = 2$ and $d = 0.5$ (Spratt, 2013). The x,y plot of the Tinkerbell map is given in Figure 4.

$$\begin{aligned} X_{n+1} &= X_n^2 - Y_n^2 + aX_n + bY_n \\ Y_{n+1} &= 2X_n Y_n + cX_n + dY_n \end{aligned} \quad (5)$$

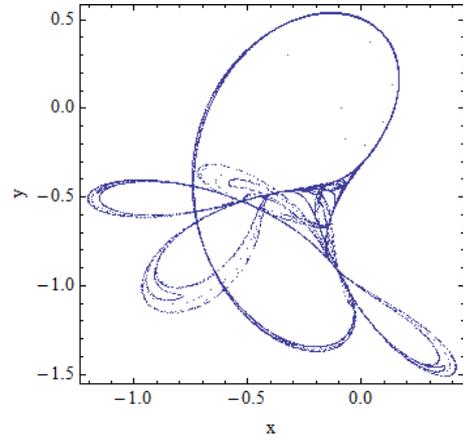


Figure 4: x,y plot of Tinkerbell map

Delayed Logistic Map

The map equations are given in (6). The control parameter $A=2.27$ (Spratt, 2013). The x,y plot of the Delayed Logistic map is given in Figure 5.

$$\begin{aligned} X_{n+1} &= AX_n(1 - Y_n) \\ Y_{n+1} &= X_n \end{aligned} \quad (6)$$

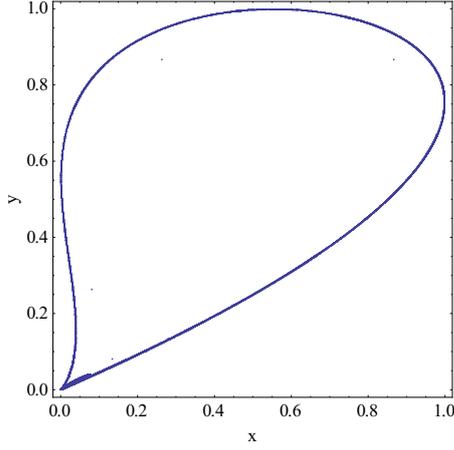


Figure 5: x, y plot of Delayed Logistic map

DIFFERENTIAL EVOLUTION, SUCCESS-HISTORY BASED ADAPTIVE DIFFERENTIAL EVOLUTION AND MULTI-CHAOTIC PARENT SELECTION

DE algorithm (Storn, Price, 1997) has four control parameters – population size NP , maximum number of generations G_{max} , crossover rate CR and scaling factor F . In the canonical form of DE, those four parameters are static and depend on the user setting. Other important features of DE algorithm are mutation strategy and crossover strategy. This work uses “rand/1/bin” mutation strategy (7) and binomial crossover (10). The success-history based adaptive differential evolution (SHADE) algorithm, on the other hand, uses only three control parameters – population size NP , maximum number of generations G_{max} and new parameter H - size of historical memories. F and CR parameters are automatically adapted based on the evolutionary process. Values of F and CR for each individual are generated according to (9) and (11) respectively. Also, the mutation strategy is different than that of canonical DE. Novel mutation strategy used in SHADE is called “current-to- $pbest/1$ ” and it is depicted in (8). The concept of basic operations in DE and SHADE algorithms is shown in following sections. For a detailed description on feature constraint correction, update of historical memories and external archive handling in SHADE see (Tanabe, Fukunaga, 2013).

Initialization

The initial population is generated randomly from objective space and has NP individuals in both algorithms. In SHADE algorithm, the external archive A is initially empty with a maximum size of NP and historical memories M_{CR} and M_F are both set to the size H where $M_{CR,i} = M_{F,i} = 0.5$ for $(i = 1, \dots, H)$.

Mutation Strategies and Parent Selection

In canonical forms of both algorithms, parent vectors are selected by classic PRNG with uniform distribution. Mutation strategy “rand/1/bin” uses three random parent vectors with indexes $r1$, $r2$ and $r3$, where $r1 = U[1, NP]$, $r2 = U[1, NP]$, $r3 = U[1, NP]$ and $r1 \neq r2 \neq r3$. Mutated vector $\mathbf{v}_{i,G}$ is obtained from three different vectors \mathbf{x}_{r1} , \mathbf{x}_{r2} , \mathbf{x}_{r3} from current generation G with help of static scaling factor $F_i = F$ as follows:

$$\mathbf{v}_{i,G} = \mathbf{x}_{r1,G} + F_i(\mathbf{x}_{r2,G} - \mathbf{x}_{r3,G}) \quad (7)$$

Contrarily, SHADEs mutation strategy “current-to- $pbest/1$ ” uses four parent vectors – current i -th vector $\mathbf{x}_{i,G}$, vector $\mathbf{x}_{pbest,G}$ randomly selected from $NP \times p$ ($p = U[p_{min}, 0.2]$, $p_{min} = 2/NP$) best vectors (in terms of objective function value) from G , randomly selected vector $\mathbf{x}_{r1,G}$ from G and randomly selected vector $\mathbf{x}_{r2,G}$ from the union of G and external archive A . Where $\mathbf{x}_{i,G} \neq \mathbf{x}_{r1,G} \neq \mathbf{x}_{r2,G}$. (8)

$$\mathbf{v}_{i,G} = \mathbf{x}_{i,G} + F_i(\mathbf{x}_{pbest,G} - \mathbf{x}_{i,G}) + F_i(\mathbf{x}_{r1,G} - \mathbf{x}_{r2,G}) \quad (8)$$

The scaling factor F_i is generated from Cauchy distribution with location parameter value of $M_{F,r}$ which is randomly selected value from scale factor historical memory, and scale parameter value of 0.1 (9).

$$F_i = C[M_{F,r}, 0.1] \quad (9)$$

Crossover and Elitism

The trial vector $\mathbf{u}_{i,G}$ which is compared with original vector $\mathbf{x}_{i,G}$ is completed by crossover operation (5) and this operation is the same for both DE and SHADE algorithms. CR_i value in DE algorithm is again static $CR_i = CR$ whereas with SHADE algorithm its value is generated from a normal distribution with a mean parameter value of $M_{CR,r}$ which is randomly selected value from crossover rate historical memory and with standard deviation value of 0.1 (10).

$$\mathbf{u}_{j,i,G} = \begin{cases} \mathbf{v}_{j,i,G} & \text{if } U[0,1] \leq CR_i \text{ or } j = j_{rand} \\ \mathbf{x}_{j,i,G} & \text{otherwise} \end{cases} \quad (10)$$

Where j_{rand} is randomly selected index of a feature, which has to be updated ($j_{rand} = U[1, D]$), D is the dimensionality of the problem. (11)

$$CR_i = N[M_{CR,r}, 0.1] \quad (11)$$

Vector which will be in next generation $G+1$ is selected by elitism. When the objective function value of trial vector $\mathbf{u}_{i,G}$ is better than that of the original vector $\mathbf{x}_{i,G}$, the trial vector will be selected for the next population and the original will be placed into the external archive A . Otherwise, the original will survive and the content of A remains unchanged (12).

$$\mathbf{x}_{i,G+1} = \begin{cases} \mathbf{u}_{i,G} & \text{if } f(\mathbf{u}_{i,G}) < f(\mathbf{x}_{i,G}) \\ \mathbf{x}_{i,G} & \text{otherwise} \end{cases} \quad (12)$$

Multi-Chaotic Parent Selection

Multi-chaotic framework for parent selection process is based on ranking selection of chaotic map based PRNGs (CPRNGs). A pool of CPRNGs $Cpool$ has to be added to the EA and each CPRNG is initialized with the same probability $pc_{init} = 1/Csize$, where $Csize$ is the size of $Cpool$. For example, for five CPRNGs $Csize = 5$ and each of them will have the probability of selection $pc_{init} = 1/5 = 0.2 = 20\%$.

For each individual vector $\mathbf{x}_{i,G}$ in generation G , the chaotic generator $CPRNG_k$ is selected from the $Cpool$ according to its probability pc_k , where k is the index of selected CPRNG. This selected generator is then used to replace classic PRNG for selection of parent vectors and if the generated trial vector succeeds in elitism, the probabilities are adjusted. There is an upper boundary for the probability of selection $pc_{max} = 0.6 = 60\%$, if the selected CPRNG reached this probability, then no adjustment takes place. Whole process is depicted in (8).

$$\text{if } f(\mathbf{u}_{i,G}) < f(\mathbf{x}_{i,G}) \quad pc_j = \begin{cases} \frac{pc_j+0.01}{1.01} & \text{if } j = k \\ \frac{pc_j}{1.01} & \text{otherwise} \end{cases} \quad (13)$$

$$\text{otherwise} \quad pc_j = pc_j$$

PROBLEM DEFINITION

The following model presents a modified open VRP with profits (see, e.g., Boussier et al, 2007 for similar problems). In order to make/test our (experimental) computations/algorithm, we consider one vehicle in the model (which corresponds to travelling salesman problem modification of VRP) that does not have to return into the initial node (depot). The basic goal is to deliver cargo from source (production facility) to multiple customers at lowest cost (with maximal profit). In our setting, not every customer must be served (if it is not profitable according to objective function). There are only few links for each node meaning the network is not complete.

The network was designed as is presented in Table 1. In Table 1 each node is given alongside with its demand and neighboring nodes and their distance. The node no. 1 is the source (production facility) therefore its demand is negative. The network is depicted in Figure 6.

The objective function (14) maximizes the total profit, i.e. the revenue minus transportation cost. Equations, or in equations alternatively, (15) - (20) present a set of constraints, where: (15) and (16) guarantee that we can neither come nor leave one node more than once and, moreover (17) guarantee that we have to come and leave every node either once or not at all; (18) sets quantities $q(i)$ from the first point of the tour; (19) means that if a node/customer is visited then the

customer must be served; (20) is a capacity constraint of the quantity $q(i)$.

Table 1: Experiment setup

No.	Demand	Neighbor No. (distance)
1	-283	2 (18.39), 6 (22.39), 10 (24.48), 15 (27.57), 19 (3.18)
2	11	1 (18.39), 3 (7.24), 7 (2.18), 8 (18.81), 10 (8.18)
3	19	2 (7.24), 19 (8.36)
4	16	6 (8.95), 9 (6.61), 13 (6.18), 20 (4.48)
5	13	9 (12.74), 10 (2.14), 11 (13.76)
6	13	1 (22.39), 4 (8.95), 9 (2.95), 10 (9.67), 14 (4.88), 17 (13.62), 18 (6.9)
7	12	2 (2.18), 9 (18.68)
8	13	2 (18.81), 13 (16.23), 14 (5.22), 18 (9.)
9	19	4 (6.61), 5 (12.74), 6 (2.95), 7 (18.68), 11 (11.14), 14 (1.99)
10	10	1 (24.48), 2 (8.18), 5 (2.14), 6 (9.67), 11 (15.1), 13 (15.28), 17 (17.8), 19 (21.63)
11	15	5 (13.76), 9 (11.14), 10 (15.1), 13 (0.76), 14 (11.59), 16 (4.31)
12	20	13 (11.15), 15 (5.16), 16 (15.07), 20 (3.36)
13	16	4 (6.18), 8 (16.23), 10 (15.28), 11 (0.76), 12 (11.15), 14 (11.01)
14	20	6 (4.88), 8 (5.22), 9 (1.99), 11 (11.59), 13 (11.01), 15 (5.2), 18 (10.25), 20 (9.6)
15	18	1 (27.57), 12 (5.16), 14 (5.2)
16	11	11 (4.31), 12 (15.07)
17	17	6 (13.62), 10 (17.8), 18 (7.47), 19 (7.93)
18	11	6 (6.9), 8 (9.), 14 (10.25), 17 (7.47)
19	12	1 (3.18), 3 (8.36), 10 (21.63), 17 (7.93)
20	17	4 (4.48), 12 (3.36), 14 (9.6)

$$\max \sum_{i_1 \in I_C, i_2 \neq i_1} d(i)M_p(i_1, i_2)p - \sum_{i_1 \in I_C} [q(i)c + 1]M_d(i, i_1)M_p(i, i_1) \quad (14)$$

$$\text{s.t.} \quad \sum_{i_1 \in I_C, i_2 \neq i_1} M_p(i_1, i_2) \leq 1, \quad \forall i_1 \in I, \quad (15)$$

$$\sum_{i_1 \in I_C, i_2 \neq i_1} M_p(i, i_2) \leq 1, \quad \forall i \in I, \quad (16)$$

$$\sum_{i_1 \in I_C, i_2 \neq i_1} M_p(i_1, i_2) = \sum_{i_1 \in I_C, i_2 \neq i_1} M_p(i_2, i_1), \quad \forall i \in I_C, \quad (17)$$

$$q(i) \leq W + [d(i) - W]M_p(1, i), \quad \forall i \in I_C, \quad (18)$$

$$\sum_{i_1 \in I} d(i)M_p(i_1, i) \leq \sum_{i_1 \in I} q(i_1)M_p(i_1, i), \quad \forall i \in I_C, \quad (19)$$

$$q(i) \leq \sum_{i_1 \in I} M_p(i_1, i)W, \quad \forall i \in I \quad (20)$$

with the decision variables:

$q(i)$: quantity delivered up to i ;

$M_p(i; i_1)$: 1 if i immediately precedes i_1 ; 0 otherwise;

the sets of indices:

I : set of all nodes in the network,

I_C : set of customers,

and parameters:

$M_d(i; i_1)$: distance matrix presenting also a cost for using a path,

W : vehicle (e.g. a lorry) capacity or maximal possible production

capacity in a production node, i.e. in $i \in I - I_C$

$d(i)$: demand in a (customer) node, $i \in I$

p : unit selling price;
 c : unit transportation cost per unit of length:

EXPERIMENT SETUP

To generate experimental network (see Figure 6) that approximates real situations, we use a network generator presented in (Pavlas et.al, 2015).

In the experiment three variants of DE were compared, the original DE rand/1/bin, SHADE and proposed MC-SHADE.

The goal for the optimizing algorithm is to find the best possible route and amount of cargo with respect to profit (14).

The maximal number of objective function evaluations Max_FEs was set to 200 000. *Cpool* contained Burgers, Dissipative, Lozi, Tinkerbell and Delayed Logistic CPRNGs. Other parameters were set as follows:

DE:
 $Dim: 20; NP: 100; G_{max}: 2\ 000; F: 0.5; CR: 0.8;$

SHADE:
 $Dim: 20; NP: 100; G_{max}: 2\ 000; H = 10;$

MC-SHADE:
 $Dim: 20; NP: 100; G_{max}: 2\ 000; H = 10;$

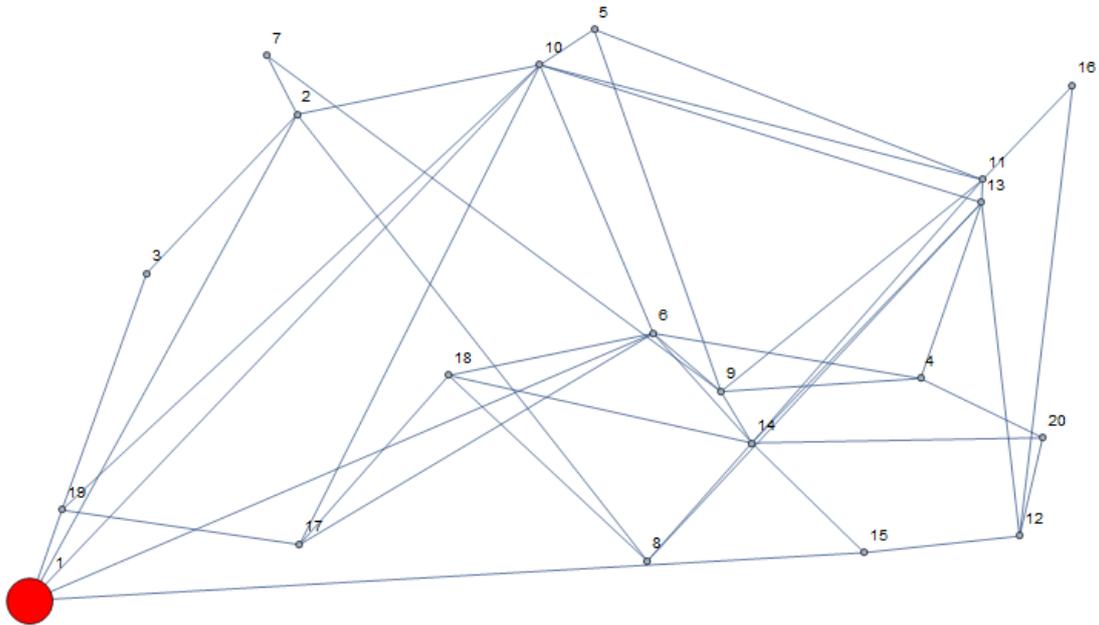


Figure 6: Experiment network setup with highlighted source of cargo

RESULTS AND DISSCUSION

In this section the results of above described experiment are presented and discussed. The statistical overview of the results is presented in Table 2. 50 independent runs were performed for each algorithm. The success rate refers to number of runs in which the best possible solution was found. The best possible solution (confirmed by deterministic commercial solver) is visualized in Figure 7.

Table 2: Results comparison

	DE	SHADE	MC-SHADE
Time for 50 runs (m:s)	1:42	8:48	9:06
Min:	13176.45	13839.60	13839.60
Max:	13865.72	13865.72	13865.72
Mean:	13588.97	13863.63	13865.20
Median:	13623.61	13865.72	13865.72
Std. Dev.:	173.35	7.16	3.69
Success Rate:	8%	92%	98%

Best solution details:

Route sequence: 1, 19, 17, 18, 6, 9, 14, 15, 12, 20, 4, 13, 11, 5, 10, 2, 3

Not-visited nodes: 7, 8, 16

Cargo picked-up in node No. 1. : 247

It is clear that the original DE is much faster than SHADE or MC-SHADE however the success rate is

unacceptable for practical use. The canonical SHADE achieved satisfactory success rate with higher time demands. With comparable time demands the proposed MC-SHADE achieved exceptional success rate failing to find the optimum only in 1 run from 50. The Wilcoxon signed-rank test between SHADE and MC-SHADE results with alternative hypothesis that mean rank value of SHADE is lower than that of MC-SHADE provided p-value of 0.0745. Based on this result the

reliability of the proposed method seems very satisfactory even in comparison to deterministic solvers. The time demand is significantly lower than those of comparable deterministic solvers where a single run for this model can easily take over 1 hour.

The presented evidence strongly supports the feasibility of the proposed multi-chaotic method for solving the modified VRP and its superiority to canonical SHADE.

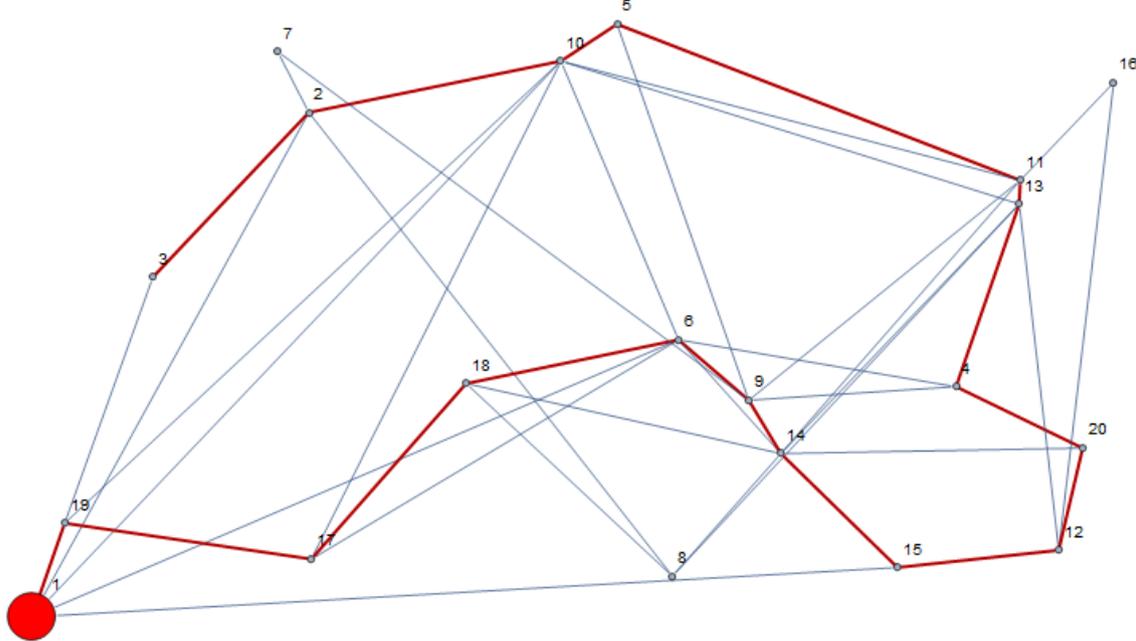


Figure 7: Final solution visualization

CONCLUSION

In this paper a Multi-chaotic Success-history based adaptive differential evolution was applied to a model of open vehicle problem with profits. The performance was compared with canonical version of the algorithm and also with the original differential evolution. The performance of proposed method is superior to both algorithms and the reliability is almost as good as a deterministic solver with significantly lower time demands. This supports the claim that the proposed method can be used as a fast and reliable transportation network problem optimizer. The future research will focus on applying these findings on similar real-world problems.

ACKNOWLEDGEMENT

This work was supported by the Programme EEA and Norway Grants for funding via grant on Institutional cooperation project nr. NF-CZ07-ICP-4-345-2016, by Grant Agency of the Czech Republic GACR P103/15/06700S, , further by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme Project no. LO1303 (MSMT-7778/2014. Also by the 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/CebiaTech/2016/007.

REFERENCES

- Avcı, M. and Topaloglu, S.: A hybrid metaheuristic algorithm for heterogeneous vehicle routing problem with simultaneous pickup and delivery. *Expert Systems with Applications*, Vol. 53, pp. 160-171, 2016
- Boussier, S., Feillet, D. and Gendreau, M.: An exact algorithm for team orienteering problems. *4OR*, Vol. 5, pp. 211-230, 2007
- Laporte, G.: The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, Vol. 59 (3), pp. 345-358, 1992
- 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 (7), pp. 1297-1304, 2011
- Pavlas, M., Nevrlý, V., Popela, P. and Somplak, R.: Heuristic for generation of waste transportation test networks. In: 21st International Conference on Soft Computing, MENDEL 2015, Brno, Czech Republic, 23-25 June, pp. 189-194, 2015
- Roupec, J., Popela, P., Hrabec, D., Novotný, J., Olstad, A., and Haugen, K.K.: Hybrid Algorithm for

- Network Design Problem with Uncertain Demands. In: Proceedings of the World Congress on Engineering and Computer Science, WCECS 2013, pp. 554-559. San Francisco, USA, 2013
- 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, pp.3087-3094, 20-23 June 2013 doi: 10.1109/CEC.2013.6557946
- Somplak, R., Prochazka, V., Pavlas, M., Popela, P.: The logistic model for decision making in waste management. Chemical Engineering Transactions, Vol. 35, pp. 817-822, 2013
- Sprott J. C.: Chaos and Time-Series Analysis. Oxford University Press, 2003
- Stodola, P., Mazal, J., Podhorec, M., and Litvaj, O.: Using the Ant Colony Optimization Algorithm for the Capacitated Vehicle Routing Problem. 16th International Conference on Mechatronics - Mechatronika (ME), pp.503-510, 2014
- Storn R., Price K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization, Vol. 11, pp.341–359, 1997
- Tanabe, R., Fukunaga, A. Success-history based parameter adaptation for differential evolution. In Evolutionary Computation (CEC), 2013 IEEE Congress on, pp. 71-78, 2013

AUTHOR BIOGRAPHIES

ADAM VIKTORIN was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the field of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@fai.utb.cz



DUSAN HRABEC is with Faculty of Applied Informatics at Tomas Bata University in Zlín. He received MSc degree in mathematical engineering at Brno University of Technology in 2011 and, recently, he works on his PhD topic in the field of applied mathematics at the same university. His research and developments efforts centre on operations research, stochastic optimization and its applications. His email address is: hrabec@fai.utb.cz



MICHAL PLUHACEK was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011 and Ph.D. in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. He now works as a researcher at the same university. His email address is: pluhacek@fai.utb.cz



STUDY ON SWARM DYNAMICS CONVERTED INTO COMPLEX NETWORK

¹Michal Pluhacek, ¹Roman Senkerik, ¹Jakub Janostik, ¹Adam Viktorin and ²Ivan Zelinka

¹Tomas Bata University in Zlin , Faculty of Applied Informatics
Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
{pluhacek,senkerik,janostik,aviktorin}@fai.utb.cz

²Department of Computer Science, Faculty of Electrical Engineering and Computer Science
VSB-TUO, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
ivan.zelinka@vsb.cz

KEYWORDS

Swarm Intelligence, Particle Swarm Optimization, Firefly Algorithm, Complex Network

ABSTRACT

In this study it is presented a summarization of our research of possible ways of creating of complex networks from the inner dynamics of Swarm Intelligence based algorithms. The particle swarm optimization algorithm and the firefly algorithm are studied in this paper. Several methods of complex network creation are proposed and discussed alongside with possibilities for future research and application.

INTRODUCTION

The Particle Swarm Optimization (PSO) (Kennedy, Eberhart 1995, Shi, Eberhart, 1998, Kennedy 1997, Nickabadi et al., 2011) and Firefly algorithm (Yang, 2008, 2009, 2013, Tilahun, 2012) are among the most prominent members of Swarm Intelligence based algorithms. These evolutionary computational techniques (ECTs) are in recent years in the center of interest of the research community. Recently the links between ECTs and complex networks (CNs) has been studied (Zelinka 2011a, 2011b, 2013).

In this study it is presented the possibilities of successful CNs creation from two swarm algorithms. Despite that the algorithms do differ the created networks seem to share similarities and in future various statistical methods may be used in order to gather information about the otherwise hidden inner dynamic of the swarm algorithms. The complex networks have many unique attributes that may help to understand and analyze the inner dynamic of Swarm algorithms. The goal is to use gathered knowledge to improve the performance of the optimization method. The usefulness of such approach was already shown in (Davendra, 2014a, 2014b).

In this study a methodology for complex network creation for PSO and Firefly Algorithm is presented. The rest of the paper is structured as follows: In the next section the PSO algorithm is described. Following is the description of Firefly Algorithm. The experimental

details alongside with methodology for CN creation and first visualizations are given in following two sections. Afterwards the conclusions are presented.

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). 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 Eq. 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 Eq. 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 Eq. 3, where t stands for current iteration number and n stands for the total number of iterations.

$$w = w_{start} - \frac{((w_{start} - w_{end}) \cdot t)}{n} \quad (3)$$

FIREFLY ALGORITHM

Firefly algorithm was first presented by Xin-She Yang in at Cambridge University (Yang, 2008, 2009). FA is based on simplified behavior of fireflies in night. Following rules were established to describe mentioned behavior (Yang, 2008, 2009, 2013, Tilahun, 2012):

1. All fireflies are unisex so that fireflies will attract each other regardless of their sex.
2. The attractiveness is proportional to the brightness, and they both decrease as their distance increases. This means that for any two flashing fireflies, the less bright one will move towards the brighter one. Firefly will move randomly if there is no brighter one.
3. The brightness of a firefly is determined by the landscape of the objective function.

Firefly's attractiveness is determined by its light intensity, which is proportional to the encoded objective function. The brightness $I(r)$ varies with the distance r monotonically and exponentially Eq. 4. That is,

$$I(r) = \frac{I_0}{1 + \lambda r^2}, \quad (4)$$

,where I_0 is the initial brightness and λ is the light absorption coefficient. Similarly, the attractiveness of a firefly can be defined using following formula Eq. 5:

$$A(r) = \frac{A_0}{1 + \lambda r^2}, \quad (5)$$

,where A_0 is the initial attractiveness.

If a firefly located at $\hat{x} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$ is brighter than firefly located at $x = (x_1, x_2, \dots, x_n)$, the firefly located at x will move towards one located at \hat{x} .

The algorithm can be summarized as follows:

1. Generate a random solution set $\{x_1, x_2, \dots, x_n\}$.

2. Compute intensity $\{I_1, I_2, \dots, I_n\}$ for each member of solution set.
3. Move firefly towards other brighter firefly if possible, move it randomly if not.
4. Update solution set.
5. If a termination criterion is fulfilled terminate algorithm. Otherwise go to step 2.

PSO EXPERIMENTS

During the experiments several different ways of complex network creation and visualization were tested. The goal was to capture the inner dynamics of swarm algorithms in a sufficient detail but in a network of appropriate size for further processing.

In the first experiment the PSO with typical defaults setting was used to optimize the commonly used Schwefel's benchmark function for 100 iterations with population size set to 30.

In this experiment the main interest was in the communications that leads to population quality improvement. Therefore only communication leading to improvement of the particles personal best (pBest) was tracked. The link was created between the particle that has improved and particle that triggered the current gBest's update.

In Figure 1 the created complex network is visualized. Nodes of similar color represent particles with same ID during different iterations. All links are from particle that triggered gBest update to particle that has improved based on that gBest.

In Figure 2 a zoomed partial view of the network is presented. It is possible to clearly see the density of the network and links of various lengths.

Close look on a single cluster in the network is presented in Figure. 3. The numbers in nodes represent a code for a particle ID and current iteration. That way it is possible to track exactly the development of the network and the communication that happens within the swarm. On this example cluster it can be observed a single gBest update led to improvement of multiple particles in different iterations.

A different visualization method was used in Figure. 4 where a smaller network is depicted. Both networks share many similarities.

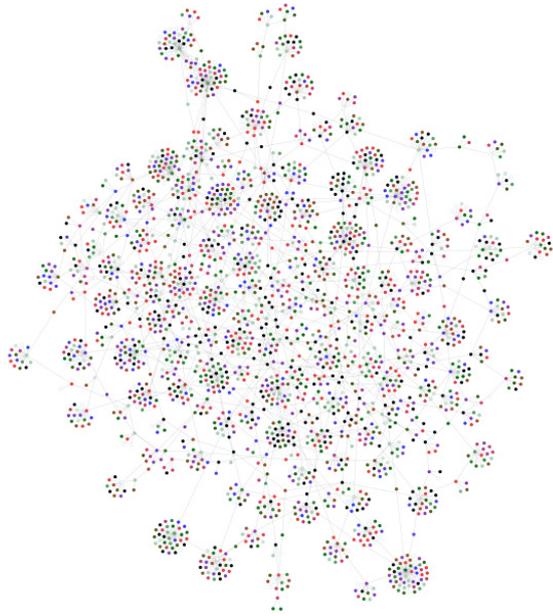


Figure 1: PSO dynamic as complex network – complete view

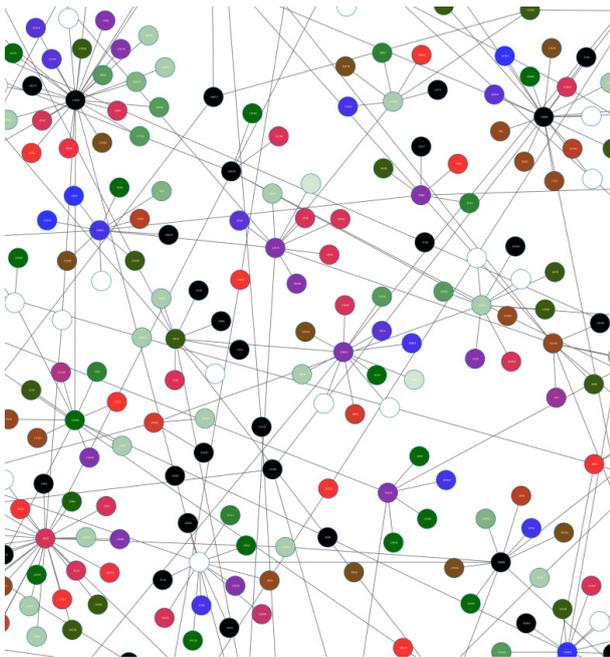


Figure 2: PSO dynamic as complex network – partial view

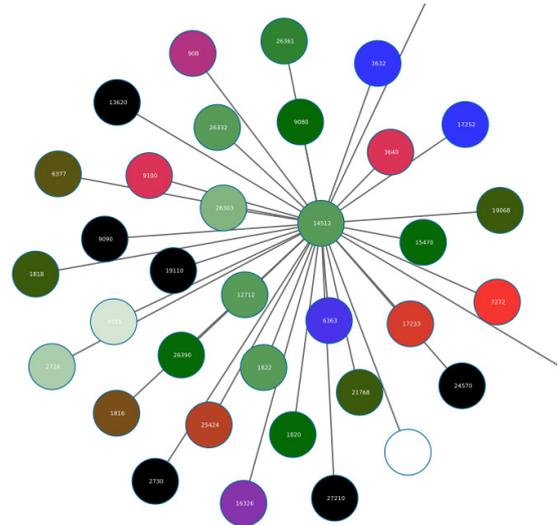


Figure 3: PSO dynamic as complex network – close view

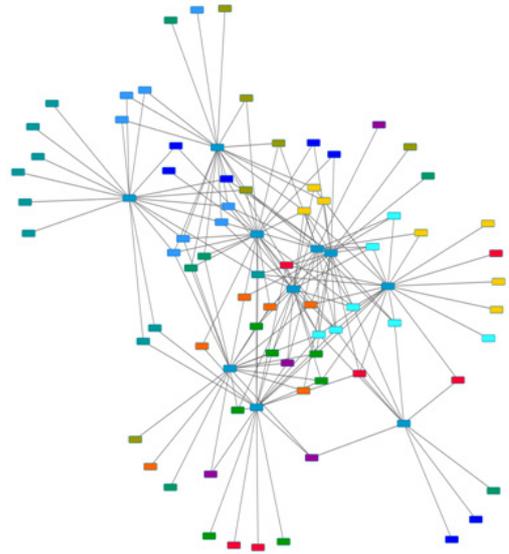


Figure 4: PSO dynamic as complex network – swarm of size 10 running for 10 iterations. Nodes with same color belong to same iteration.

FIREFLY EXPERIMENTS

In the second experiment the firefly algorithm was used. The algorithm optimized Schwefel's benchmark function for 100 iterations with population of size 30. In the process of creation of the network every firefly was visualized as a node. Connection between nodes is plotted for every successful interaction between fireflies. Successful interaction is defined as such interaction where one of the individuals gets improved. In the case of FA it is when firefly flies towards another and improves own brightness. This leads to network presented in Figure 5 and Figure 6. Duplicate connections were omitted from the network in the sake of clarity.

Because across multiple iterations of algorithm there may be multiple connections between nodes the connections were weighted in this design. If there is connection between the firefly A and B it starts with weight 1. If in another iteration there is another successful interaction between the firefly A and B, new connection is not created but the weight of the existing connection is incremented by 1. At the end of evolution the weight is normalized. If the firefly gets improved in all iterations, at the end of the evolution their connection will have weight 1. If it never gets improved their connection will have weight 0.

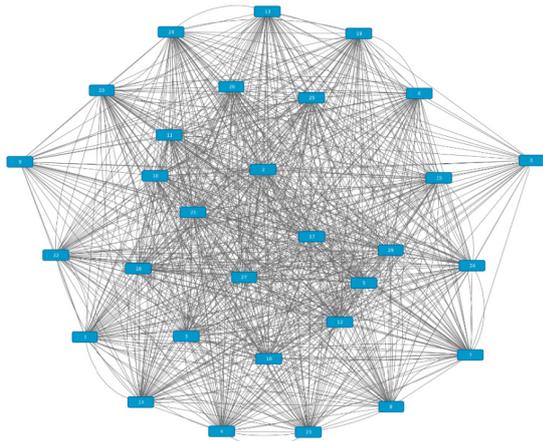


Figure 5: Basic weighted oriented network for population of size 30 after 100 iterations.

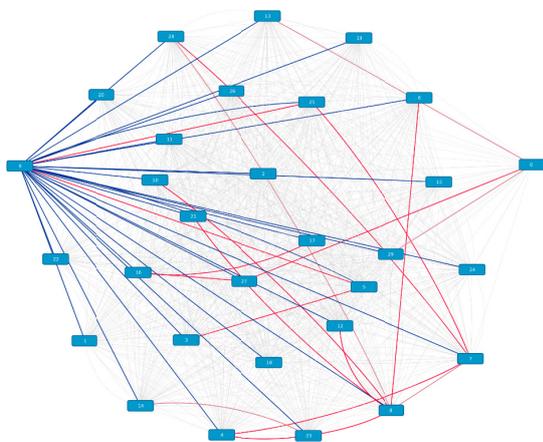


Figure 6: Basic weighted oriented network for population of size 30 after 100 iterations with visually highlighted weights (Blue: $0.7 < \text{weight}$, Grey: $0.3 < \text{weight} < 0.7$, Red: $\text{weight} < 0.3$).

Finally, even though previous network yield interesting information, its size is limited by the size of population. This may not be favorable for all tools of network analysis. In the extended model (Figure 7 – 9) nodes are not created by fireflies alone but by fireflies and the iteration in which they were created. This way every firefly will get represented up to number of times that is equal to the number of iterations. Because of this there cannot be multiple connections between same nodes, so the weights become unnecessary.

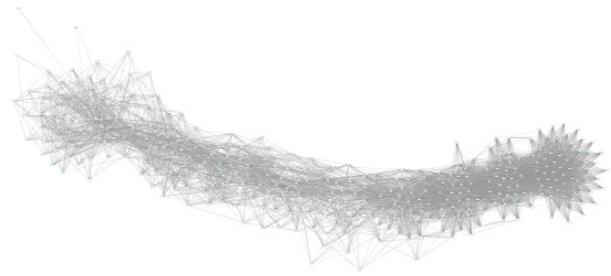


Figure 7: Time capturing oriented network

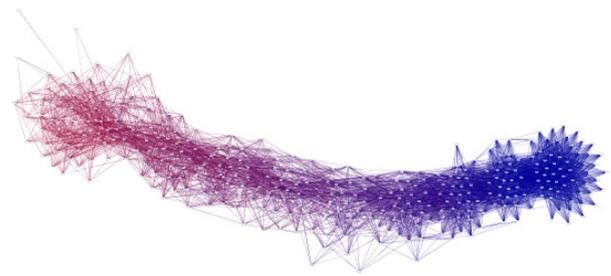


Figure 8: Time capturing oriented network with visually encoded iteration order (Red: iteration 1, Blue: iteration 10)

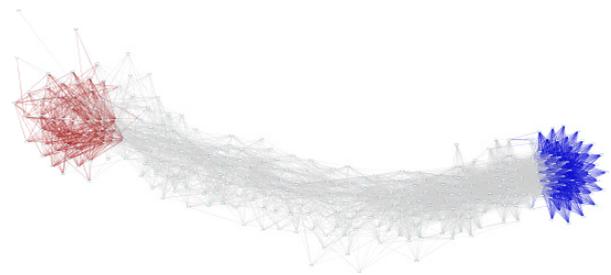


Figure 9: Time capturing oriented network with highlighted iteration 0 and 10.

CONCLUSION

In this study several complex networks were created from Swarm Intelligence based algorithms and analyzed. The goal is to capture the hidden inner dynamics of swarm algorithms. The information based on the complex network analysis may be used in various adaptive approaches. The complex networks may prove a very beneficial tool for capturing the inner dynamics

of swarm algorithms. The future research will shift the focus from creating the networks to implement various analytic and statistical tools and further to impellent adaptive mechanisms in order to improve the performance of the optimization algorithm

ACKNOWLEDGEMENT

This work was supported by Grant Agency of the Czech Republic – GACR P103/15/06700S, further by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and also by the 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/CebiaTech/2016/007.

REFERENCES

- Davendra, D., Zelinka, I., Metlicka, M., Senkerik, R., Pluhacek, M., "Complex network analysis of differential evolution algorithm applied to flowshop with no-wait problem," *Differential Evolution (SDE)*, 2014 IEEE Symposium on , vol., no., pp.1,8, 9-12 Dec. 2014
- Davendra, D., Zelinka, I., Senkerik, R. and Pluhacek, M. Complex Network Analysis of Discrete Self-organising Migrating Algorithm, in: Zelinka, I. and Suganthan, P. and Chen, G. and Snasel, V. and Abraham, A. and Rossler, O. (Eds.) *Nostradamus 2014: Prediction, Modeling and Analysis of Complex Systems, Advances in Intelligent Systems and Computing*, Springer Berlin Heidelberg, pp. 161–174 (2014).
- Kennedy J. and Eberhart R., "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, 1995, pp. 1942–1948.
- Kennedy J., "The particle swarm: social adaptation of knowledge," in *Proceedings of the IEEE International Conference on Evolutionary Computation*, 1997, pp. 303–308."
- Nickabadi A., Ebadzadeh M. M., Safabakhsh R., 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
- Shi Y. and Eberhart R., "A modified particle swarm optimizer," in *Proceedings of the IEEE International Conference on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, 1998, pp. 69–73.I. S.
- Tilahun, Surafel Lulseged; ONG, Hong Choon. Modified Firefly Algorithm. *Journal of Applied Mathematics*. 2012, vol. 2012. ISSN:1110-757X.
- Yang X. S., *Nature-Inspired Metaheuristic Algorithms*, Luniver Press, UK, (2008).
- Yang X. S., Firefly algorithms for multimodal optimization, *Proc. 5th Symposium on Stochastic Algorithms, Foundations and Applications*, (Eds. O. Watanabe and T. Zeugmann), *Lecture Notes in Computer Science*, 5792: 169-178 (2009).
- Yang X. S.; Xingshi H. *Firefly algorithm: Recent Advances and Applications*. 2013.
- Zelinka, I. Investigation on relationship between complex network and evolutionary algorithms dynamics, *AIP Conference Proceedings* 1389 (1) 1011–1014 2011a.
- Zelinka, I., Davendra, D., Enkek, R., Jaek, R.: Do Evolutionary Algorithm Dynamics Create Complex Network Structures? *Complex Systems* 2, 0891–2513, 20, 127–140, 2011b
- Zelinka, I., Davendra, D.D., Chadli, M., Senkerik, R., Dao, T.T., Skanderova, L.: Evolutionary Dynamics as The Structure of Complex Networks. In: Zelinka, I., Snasel, V., Abraham, A. (eds.) *Handbook of Optimization*. ISRL, vol. 38, pp. 215–243. Springer, Heidelberg (2013)

AUTHOR BIOGRAPHIES

MICHAL PLUHACEK was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Information Technologies and obtained his MSc degree in 2011 and Ph.D. in 2016 with the dissertation topic: Modern method of development and modifications of evolutionary computational techniques. He now works as a researcher 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 Zlín, 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



ADAM VIKTORIN was born in the Czech Republic, and went to the Faculty of Applied Informatics at Tomas Bata University in Zlín, where he studied Computer and Communication Systems and obtained his MSc degree in 2015. He is studying his Ph.D. at the same university and the field of his studies are: Artificial intelligence, data mining and evolutionary algorithms. His email address is: aviktorin@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 Zlín. Now he is a professor at Technical University of Ostrava (research in: Artificial Intelligence, Theory of Information). Email address: ivan.zelinka@vsb.cz



ON THE SIMULATION OF COMPLEX CHAOTIC DYNAMICS FOR CHAOS BASED OPTIMIZATION

¹Roman Senkerik, ¹Michal Pluhacek, ¹Adam Viktorin, 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

KEYWORDS

Deterministic chaos; Chaotic oscillators; Heuristic; Chaotic Optimization; Chaotic Pseudo Random Number Generators

ABSTRACT

This paper investigates the utilization of the complex chaotic dynamics given by the several selected time-continuous chaotic systems, as the chaotic pseudo random number generators and driving maps for the chaotic optimization. Chaos based optimization concept is utilizing direct output iterations of chaotic system transferred into the required numerical range or it uses the chaotic dynamics for mapping the search space mostly within the local search techniques. This paper shows two groups of complex chaotic dynamics given by either chaotic flows or oscillators. Simulations of examples of chaotic dynamics transferred to the pseudo random number generators were performed and related issues like distributions of such a systems, periodicity, and dependency on sampling times are briefly discussed in this paper.

INTRODUCTION

Generally speaking, the term “chaos” can denote anything that cannot be predicted deterministically. In the case that the word “chaos” is combined with an attribute such as “deterministic”, then a specific type of chaotic phenomena is involved, having their specific laws, mathematical apparatus and a physical origin. The deterministic chaos is a phenomenon that - as its name suggests - is not based on the presence of a random or any stochastic effects. It is clear from the structure of the equations (see the section *Chaotic Optimization*), that no mathematical term expressing randomness is present. The seeming randomness in deterministic chaos is related to the extreme sensitivity to the initial conditions (Celikovskiy and Zelinka 2010).

In the past, the chaos has been observed in many of various systems (including evolutionary one). Systems exhibiting deterministic chaos include, for instance, weather, biological systems, many electronic circuits (Chua’s circuit), mechanical systems, such as double pendulum, magnetic pendulum, or so called billiard problem.

The idea of using chaotic systems instead of random processes (pseudo-number generators - PRNGs) has

been presented in several research fields and in many applications with promising results (Lee and Chang 1996; Wu and Wang, 1999).

Another research joining deterministic chaos and pseudorandom number generator has been done for example in (Lozi 2012). Possibility of generation of random or pseudorandom numbers by use of the ultra weak multidimensional coupling of p 1-dimensional dynamical systems is discussed there.

Another paper (Persohn and Povinelli 2012), deeply investigate logistic map as a possible pseudorandom number generator and is compared with contemporary pseudo-random number generators. A comparison of logistic map results is made with conventional methods of generating pseudorandom numbers. The approach used to determine the number, delay, and period of the orbits of the logistic map at varying degrees of precision (3 to 23 bits). Another paper (Wang and Qin 2012) proposed an algorithm of generating pseudorandom number generator, which is called (couple map lattice based on discrete chaotic iteration) and combine the couple map lattice and chaotic iteration. Authors also tested this algorithm in NIST 800-22 statistical test suits and for future utilization in image encryption. In (Narendra et al. 2010) authors exploit interesting properties of chaotic systems to design a random bit generator, called CCCBG, in which two chaotic systems are cross-coupled with each other. A new binary stream-cipher algorithm based on dual one-dimensional chaotic maps is proposed in (Yang and Wang 2012) with statistic proprieties showing that the sequence is of high randomness. Similar studies are also done in (Bucolo et al. 2002).

MOTIVATION

Recently the chaos has been used also to replace pseudo-number generators (PRNGs) in evolutionary algorithms (EAs) or for mapping of solutions or iterations within local search techniques. An evolutionary chaotic approach generally uses the chaotic system in the place of a pseudo random number generator (Aydin et al. 2010). This causes the heuristic to map unique regions, since the chaotic system iterates to new regions. The task is then to select a very good chaotic system (either discrete or time-continuous) 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 in the form of chaos pseudo random number generator (CPRNG) for DE (Differential Evolution) and SOMA (Zelinka 2004) in the task of optimal PID tuning. Several papers have been recently focused on the connection of heuristic and chaotic dynamics either in the form of hybridizing of DE with chaotic searching algorithm (Liang et al. 2011) or in the form of chaotic mutation factor and dynamically changing weighting and crossover factor in self-adaptive chaos differential evolution (SACDE) (Zhenyu et al. 2006). Also the PSO (Particle Swarm Optimization) algorithm with elements of chaos was introduced as CPSO (Coelho and Mariani 2009).

This idea was later extended with the successful experiments with chaos driven DE (ChaosDE) (Senkerik et al. 2014) with both complex and simple test functions.

At the same time the chaos embedded PSO with inertia weigh strategy was closely investigated (Pluhacek et al. 2013), followed by and novel chaotic Multiple Choice PSO strategy (Chaos MC-PSO) (Pluhacek et al. 2014).

The primary aim of this work is to try, test, analyze and compare the implementation of different natural chaotic dynamic as the CPRNG, thus to analyze and highlight the different influences to the system, which utilizes the selected CPRNG (including the evolutionary computational techniques).

CHAOTIC OPTIMIZATION

It exist two possible utilizations of chaotic dynamics in optimization tasks. Either direct output simulation iterations of chaotic system are transferred into the required numerical range (as simple CPRNG) or it uses the complexity of chaotic systems for dynamical mapping of the search space mostly within the local search techniques (Hamaizia and Lozi 2011). The general idea of CPRNG is to replace the default system PRNG with the chaotic system. As the chaotic system is a set of equations with a static start position, we created a random start position of the system, in order to have different start position for different experiments. Once the start position of the chaotic system has been obtained, the system generates the next sequence using its current position.

Generally there exist many other approaches as to how to deal with the negative numbers as well as with the scaling of the wide range of the numbers given by the chaotic systems into the typical range 0 – 1:

CHAOTIC SYSTEMS

This section contains the description of time-continuous chaotic systems (flows or oscillators), which were used as the CPRNG. In this research, direct sampled output iterations of the chaotic systems were used for the generation of real numbers scaled into the typical range for random function: <0 - 1>. Following chaotic systems were used: Lorenz system (1). Rossler system

(2); unmodified UEDA oscillator (3); and Driven Van der Pol Oscillator (4).

The parametric plots of the chaotic systems are depicted in Figures 1, 4, 7 and 10. The typical chaotic behavior of the utilized chaotic systems, represented by the examples of direct outputs for the variable x is depicted in Figures 2, 5, 8 and 11. Finally the Figures 3, 6, 9 and 12 show the example of dynamical sequencing during the generating of pseudo number numbers transferred into the range <0 - 1> by means of particular studied CPRNGs and with the sampling rate of 0.5s.

Lorenz System

The Lorenz system is a 3-dimensional dynamical flow, which exhibits chaotic behavior. It was introduced by Edward Lorenz in 1963, who derived it from the simplified equations of convection rolls arising in the equations of the atmosphere. It is a very simple model of the dynamics of a fluid heated from below in the gravitation field, thus it was first used to study a problem of weather predictability. The equations which describe the Lorenz system are given in (1) (Sprott 2003). The Lorenz attractor is depicted on Fig. 1.

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x) \\ \frac{dy}{dt} &= x(\rho - z) - y \\ \frac{dz}{dt} &= xy - \beta z \end{aligned} \quad (1)$$

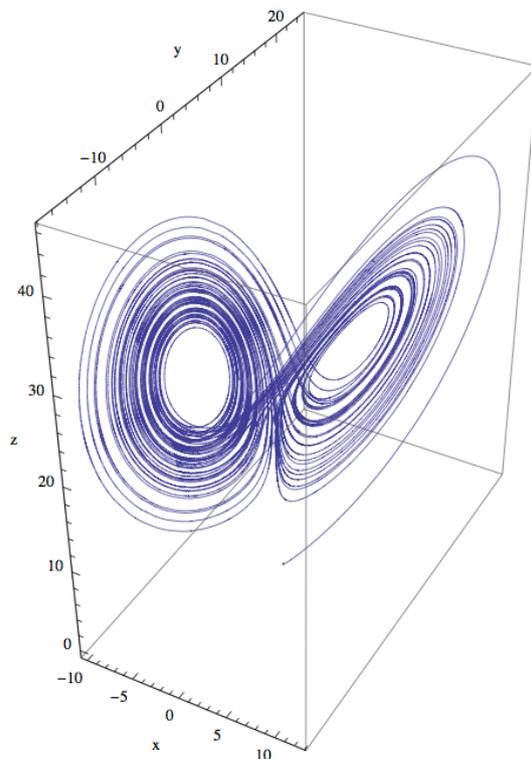


Fig. 1 x, y, z parametric plot of the Lorenz system

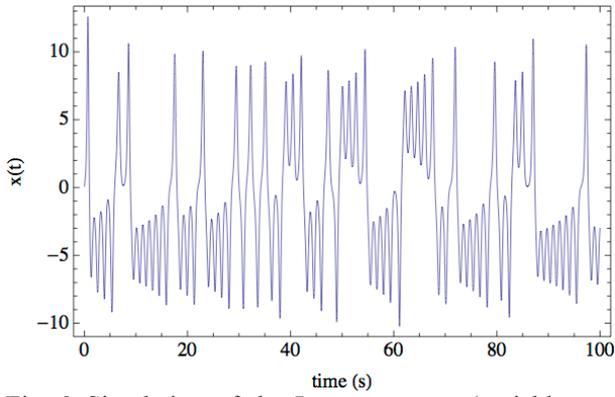


Fig. 2 Simulation of the Lorenz system (variable x – line-plot)

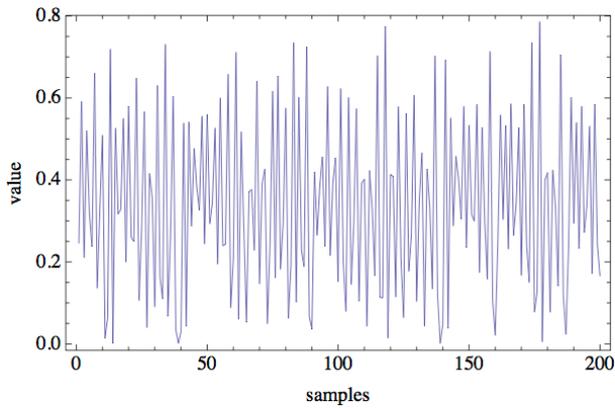


Fig. 3 Example of the chaotic dynamics: range $\langle 0 - 1 \rangle$ generated by means of the Lorenz system

Rössler System

The Rössler system is a system that exhibits chaotic dynamics associated with the fractal properties of the attractor. It was originally introduced as an example of very simple chaotic flow containing chaos similarly to the Lorenz attractor. This attractor has some similarities to the Lorenz attractor, but is simpler and has only one manifold. The attractor was later found to be useful in modeling equilibrium in chemical reactions. The Rössler system is given by following set of equations (2) (Sprott 2003):

$$\begin{aligned} \frac{dx}{dt} &= -y - z \\ \frac{dy}{dt} &= x + ay \\ \frac{dz}{dt} &= b + z(x - c) \end{aligned} \quad (2)$$

Rössler studied the chaotic attractor with $a = 0.2$, $b = 0.2$, and $c = 5.7$,

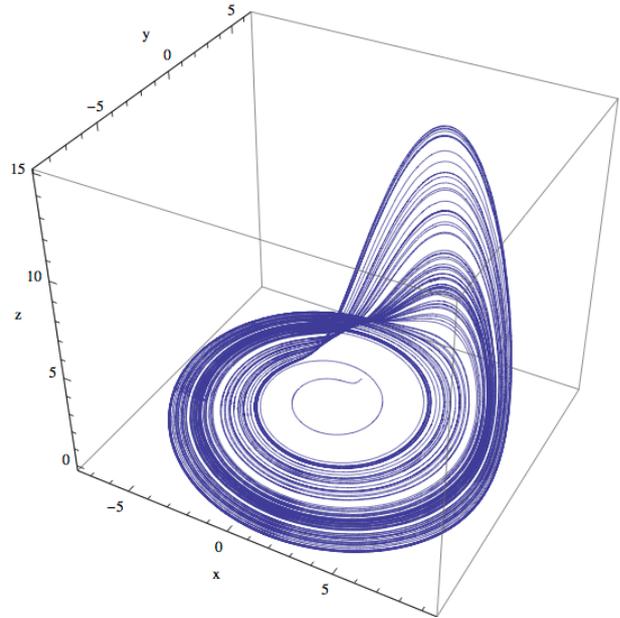


Fig. 4 x, y, z parametric plot of the Rössler system

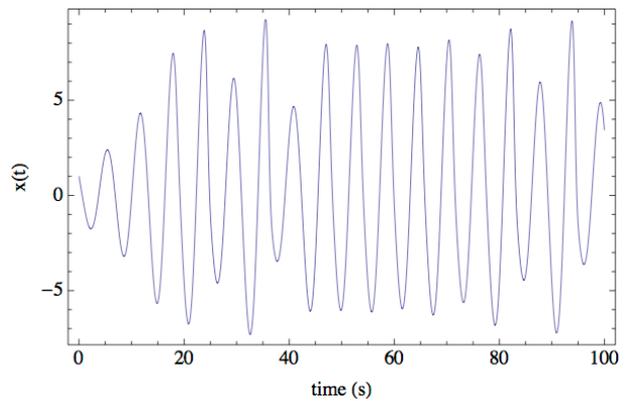


Fig. 5 Simulation of the Rössler system (variable x – line-plot)

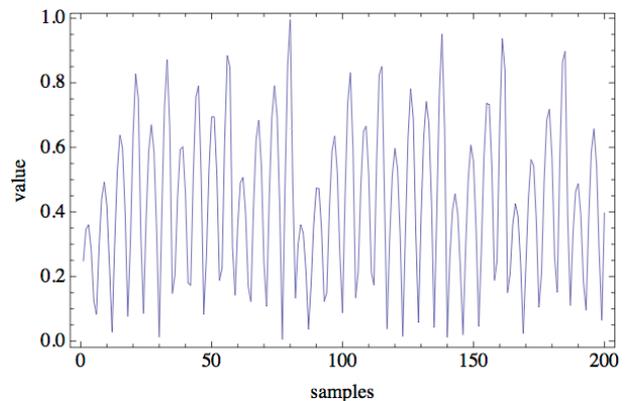


Fig. 6 Example of the chaotic dynamics: range $\langle 0 - 1 \rangle$ generated by means of the Rössler system

UEDA Oscillator

UEDA oscillator is the simple example of driven pendulums, which represent some of the most significant examples of chaos and regularity.

The UEDA system can be simply considered as a special case of intensively studied Duffing oscillator that has both a linear and cubic restoring force. Ueda oscillator represents the both biologically and physically important dynamical model exhibiting chaotic motion. It can be used to explore much physical behavior in biological systems (Sprott 2003).

The UEDA chaotic system equations are given in (3). The parameters are: $a = 1.0$, $b = 0.05$, $c = 7.5$ and $\omega = 1.0$ as suggested in (Sprott 2003). The x , y parametric plot of the chaotic system is depicted in Fig. 7.

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -ax^3 - by + c \sin \omega t \end{aligned} \quad (3)$$

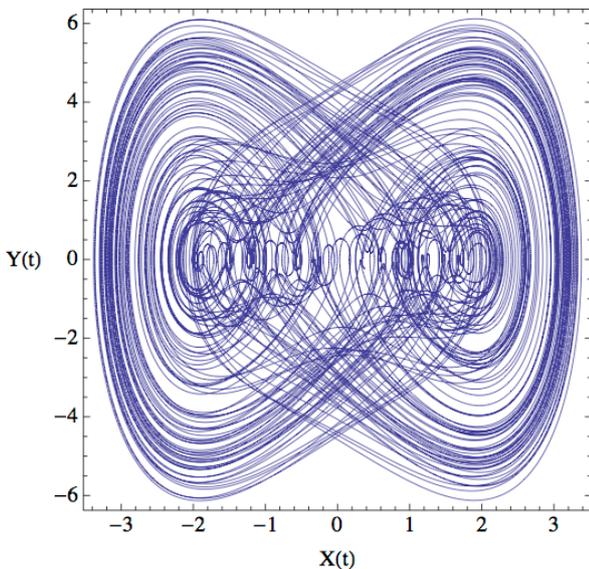


Fig. 7 x , y parametric plot of the UEDA oscillator

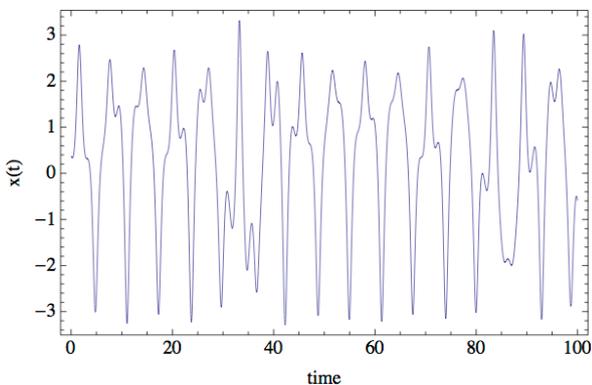


Fig. 8 Simulation of the UEDA oscillator (variable x – line-plot)

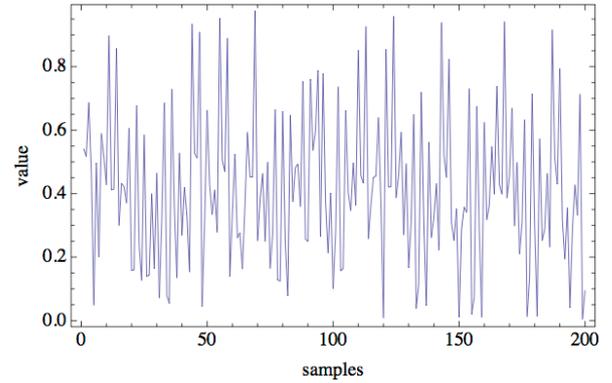


Fig. 9 Example of the chaotic dynamics: range $<0 - 1>$ generated by means of the UEDA chaotic oscillator

Van der Pol Oscillator

Van der Pol oscillator is the simple example of the limit cycles and chaotic behavior in electrical circuits employing vacuum tubes. Similarly to the UEDA oscillator, it can be used to explore physical (unstable) behaviour in biological sciences. (Bharti and Yuasa 2010).

In this paper, the forced, or commonly known as driven, Van der Pol oscillator is investigated. This system consist of the original Van der Pol oscillator definition with the added driving function $a \sin(\omega t)$, thus the differential equations have the form (4). The parameters are: $\mu = 0.2$, $\gamma = 8.0$, $a = 0.35$ and $\omega = 1.02$ as suggested in (Sprott 2003).

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= \mu(1 - \gamma x^2)y - x^3 + a \sin \omega t \end{aligned} \quad (4)$$

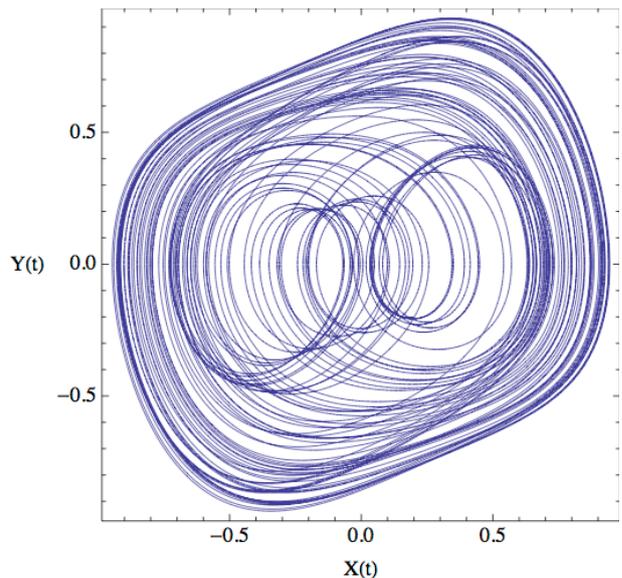


Fig.10 x , y parametric plot of the Van der Pol oscillator

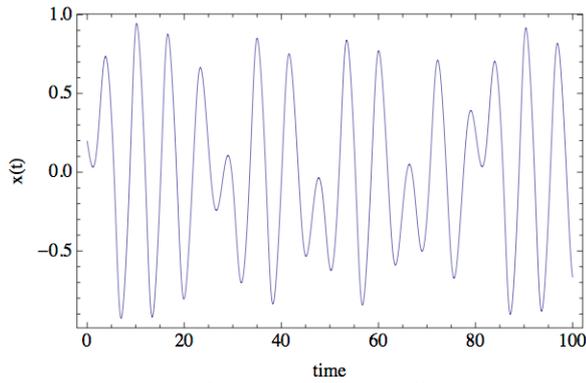


Fig. 11 Simulation of the Van der Pol oscillator (variable x – line-plot)

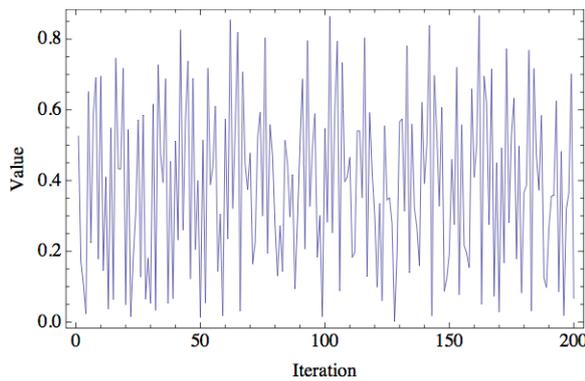


Fig. 12 Example of the chaotic dynamics: range $<0 - 1>$ generated by means of the Van der Pol chaotic oscillator

EXPERIMENT DISCUSSION

Four different complex chaotic time-continuous flows/oscillators have been simulated and the output behaviors have been transferred into the pseudo random

number sequences. Comparisons of the distributions of pseudo random real numbers transferred into the range $<0 - 1>$ are depicted in Fig.13. Findings can be summarized as follows:

- It was proven that chaos based optimization is very sensitive to the hidden chaotic dynamics driving the CPRNG/mapping the search space. Such a chaotic dynamics can be significantly changed by the selection of sampling time in the case of the time-continuous systems. Only small sampling rate of 0.1s keeps the information about the chaotic dynamics (as in Figs 14 and 15) and by using such chaotic dynamics driving the optimization technique, its performance is significantly influenced.
- Oscillators are giving more dynamical pseudo random sequences with unique quasi-periodical sequencing in comparison with chaotic flows (See Figs 3, 6, 9 and 12).
- Furthermore presented chaotic systems have additional accessible 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 Chaos based optimization (including adaptive switching between chaotic systems or sampling rates).
- Distributions are similar for all systems.

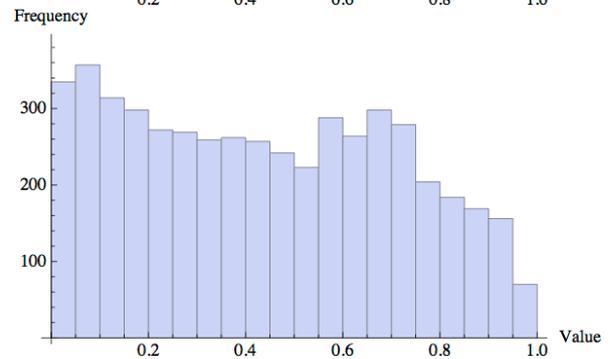
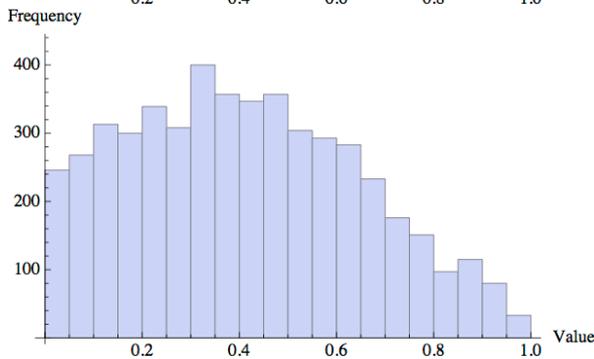
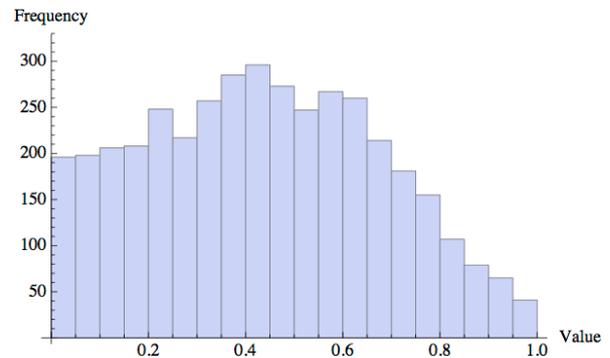
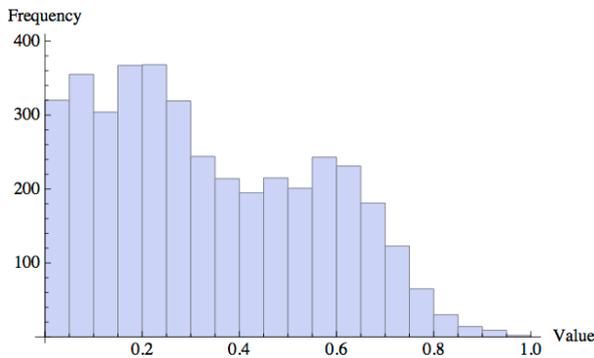


Fig. 13: Comparison of distributions of pseudo random real numbers transferred into the range $<0 - 1>$ (5000 samples); upper left – Lorenz system, upper right Rossler system, bottom left UEDA oscillator, bottom right Van der Pol Oscillator.

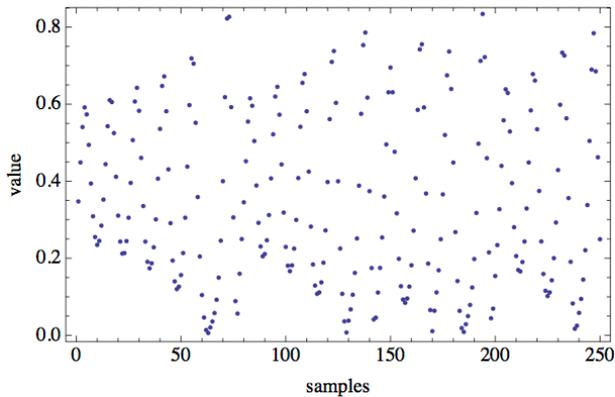


Fig. 14 Example of the chaotic dynamics: range $<0 - 1>$ generated by the UEDA oscillator; sampling rate 0.1s. The chaotic dynamics is kept in the pseudo random sequence.

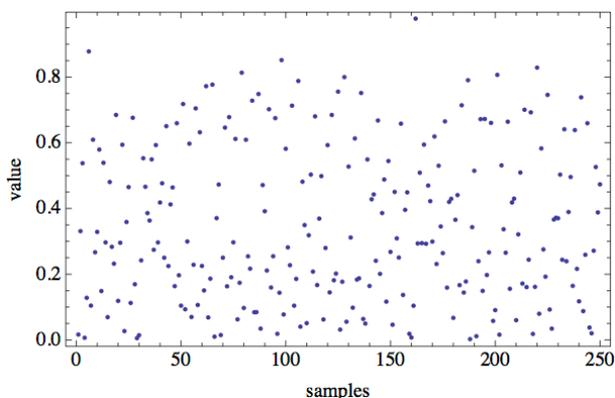


Fig. 15 Example of the chaotic dynamics: range $<0 - 1>$ generated by the UEDA oscillator; sampling rate 0.5s. The chaotic dynamics is more suppressed and averaged in the pseudo random sequence.

CONCLUSION

The novelty of this research represents investigation on the utilization of the complex chaotic dynamics given by the several selected time-continuous chaotic systems, as the chaotic pseudo random number generators and driving maps for the chaos based optimization. This paper showed two groups of complex chaotic dynamics given by either chaotic flows or oscillators.

Future plans are including the testing of combination of different time-continuous chaotic systems as well as the adaptive switching between systems, adaptive or chaos-like sequencing sampling rates and obtaining a large number of results to perform statistical tests.

ACKNOWLEDGEMENT

This work was supported by Grant Agency of the Czech Republic - GACR P103/15/06700S, further by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and also by the 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/CebiaTech/2016/007.

REFERENCES

- Aydin, I., Karakose, M. and Akin, E. (2010) 'Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection', *Expert Systems with Applications*, 37(7), 5285-5294.
- Bharti, L. and Yuasa, (2010) M. Energy Variability and Chaos in Ueda Oscillator. Available: <http://www.rist.kindai.ac.jp/no.23/yuasa-EVCUO.pdf>
- Bucolo, M., Caponetto, R., Fortuna, L., Frasca, M. and Rizzo, A. (2002) 'Does chaos work better than noise?', *Circuits and Systems Magazine, IEEE*, 2(3), 4-19.
- Caponetto, R., Fortuna, L., Fazzino, S. and Xibilia, M. G. (2003) 'Chaotic sequences to improve the performance of evolutionary algorithms', *IEEE Transactions on Evolutionary Computation*, 7(3), 289-304.
- Celikovsky, S. and Zelinka, I. (2010) 'Chaos Theory for Evolutionary Algorithms Researchers' in Zelinka, I., Celikovsky, S., Richter, H. and Chen, G., eds., *Evolutionary Algorithms and Chaotic Systems*, Springer Berlin Heidelberg, 89-143.
- 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*, 39(2), 510-518.
- Davendra, D., Zelinka, I. and Senkerik, R. (2010) 'Chaos driven evolutionary algorithms for the task of PID control', *Computers & Mathematics with Applications*, 60(4), 1088-1104.
- Hamaizia, T. and Lozi, R. (2011) *Improving Chaotic Optimization Algorithm using a new global locally averaged strategy*, translated by pp. 17-20.
- Lee, J. S. and Chang, K. S. (1996) 'Applications of chaos and fractals in process systems engineering', *Journal of Process Control*, 6(2-3), 71-87.
- Liang, W., Zhang, L. and Wang, M. (2011) 'The chaos differential evolution optimization algorithm and its application to support vector regression machine', *Journal of Software*, 6(7), 1297-1304.
- Lozi, R. (2012) 'Emergence of Randomness from Chaos', *International Journal of Bifurcation and Chaos*, 22(02), 1250021.
- Narendra, K. P., Vinod, P. and Krishan, K. S. (2010) 'A Random Bit Generator Using Chaotic Maps', *International Journal of Network Security*, 10, 32 - 38.
- Persohn, K. J. and Povinelli, R. J. (2012) 'Analyzing logistic map pseudorandom number generators for periodicity induced by finite precision floating-point representation', *Chaos, Solitons & Fractals*, 45(3), 238-245.
- Pluhacek, M., Senkerik, R. and Zelinka, I. (2014) 'Multiple Choice Strategy Based PSO Algorithm with Chaotic Decision Making – A Preliminary Study' in Herrero, Á., Baruque, B., Klett, F., Abraham, A., Snášel, V., Carvalho, A. C. P. L. F., Bringas, P. G., Zelinka, I., Quintián, H. and Corchado, E., eds., *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, Springer International Publishing, 21-30.
- Pluhacek, M., Senkerik, R., Davendra, D., Kominkova Oplatkova, Z. and Zelinka, I. (2013) 'On the behavior and performance of chaos driven PSO algorithm with inertia weight', *Computers & Mathematics with Applications*, 66(2), 122-134.
- 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 Herrero, Á., Baroque, B., Klett, F., Abraham, A., Snášel, V., Carvalho, A. C. P. L. F., Bringas, P. G., Zelinka, I., Quintián, H. and Corchado, E., eds., *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, Springer International Publishing, 41-50.
- Sprott, J. C. (2003) *Chaos and Time-Series Analysis*, Oxford University Press.
- Wang, X.-y. and Qin, X. (2012) 'A new pseudo-random number generator based on CML and chaotic iteration', *Nonlinear Dynamics*, 70(2), 1589-1592.
- Wu, J., Lu, J. and Wang, J. (2009) 'Application of chaos and fractal models to water quality time series prediction', *Environmental Modelling & Software*, 24(5), 632-636.
- Yang, L. and Wang, X.-Y. (2012) 'Design of Pseudo-random Bit Generator Based on Chaotic Maps', *International Journal of Modern Physics B*, 26(32), 1250208.
- Zelinka, I. (2004) 'SOMA — Self-Organizing Migrating Algorithm' in *New Optimization Techniques in Engineering*, Springer Berlin Heidelberg, 167-217.
- Zhenyu, G., Bo, C., Min, Y. and Binggang, C. (2006) 'Self-Adaptive Chaos Differential Evolution' in Jiao, L., Wang, L., Gao, X.-b., Liu, J. and Wu, F., eds., *Advances in Natural Computation*, Springer Berlin Heidelberg, 972-975.

SIMULATION OF SUBMARINE GROUNDWATER DISCHARGE OF DISSOLVED ORGANIC MATTER USING CELLULAR AUTOMATA

Lars Nolle, Holger Thormählen and Harald Musa
Department of Engineering Science
Jade University of Applied Science
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
Email: lars.nolle@jade-hs.de

KEYWORDS

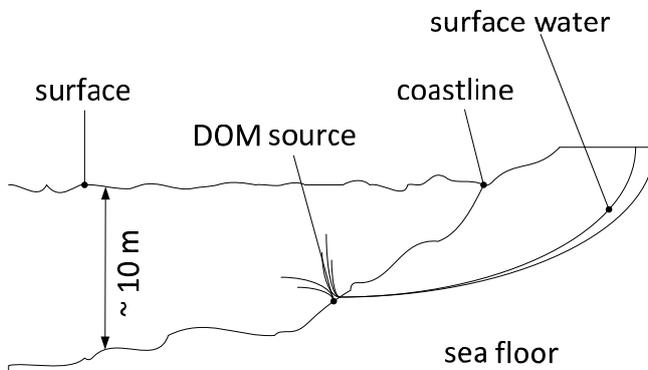
Dissolved organic matter, costal water simulation, cellular automata.

ABSTRACT

In order to design new search strategies for collaborating autonomous underwater vehicles, a 2D simulator was developed to simulate costal water environments. This allows for evaluating the new strategies without running the risk of losing expensive hardware during the tests. The simulator developed is based on the concept of cellular automata. Details of the simulator are described before preliminary qualitative results for three different scenarios are discussed. Finally, planned further improvements are presented.

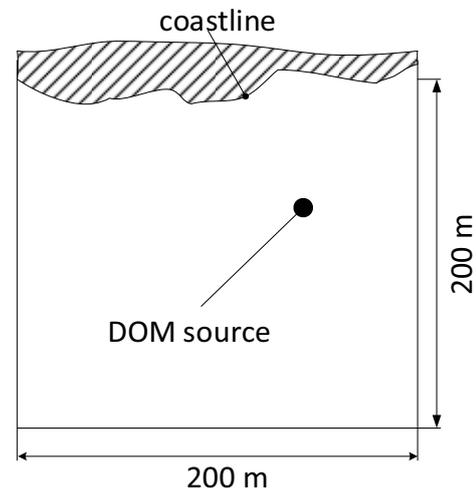
INTRODUCTION

Marine scientists are sometimes interested in locating submarine underground discharges of dissolved organic matter (DOM) (Suryaputra 2012) near the coastline. The major source of DOM in coastal waters is the degradation of terrestrial plant matter, which is dissolved and, for example, transported through river systems and estuaries to the marine environment (Stedmon et al. 2003). DOM sources or springs may appear if surface water transporting DOM percolates through the sea floor and exits beneath sea level (Figure 1).



Figures 1: Submarine Groundwater Discharge of Dissolved Organic Matter Near the Coastline

The area of interest, i.e. the search area, is reasonable small. Figure 2 shows the top view of a typical scenario. The long term of this project is to develop a swarm of small autonomous underwater vehicles (AUVs) (Figure 3) to locate any DOM sources, i.e. locations of highest DOM concentration, within a predefined area of interest (Nolle 2015).



Figures 2: Top View of the Area of Interest

In the first stage of the project, suitable search strategies for the coordinated operation of collaborating AUVs are to be developed and successively evaluated.



Figures 3: OpenROV to be Modified for Autonomous Operation

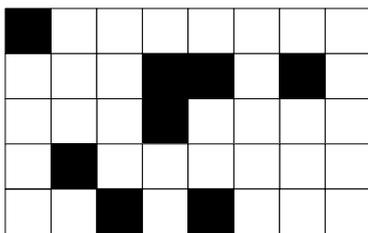
In order to be able to test different collaboration schemes for AUVs, without the risk of losing expensive hardware during the trials, a simulation of the marine environment is required. The next section describes the simulation of submarine groundwater discharges of DOM using cellular automata.

SIMULATION

Since the simulation developed is based on cellular automata, this section introduces briefly the concept of cellular automata before providing details about the simulation tool developed.

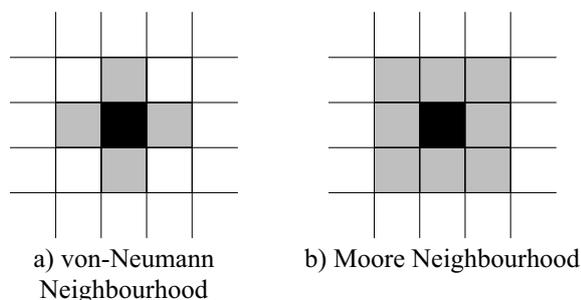
Cellular Automata

Cellular automata (CA) are general dynamic models of complex systems, which are discrete in time and space (Wolfram 1984). They consist of ordered collections of simple identical cells. Each cell has a current (discrete) state, which is updated after every time step (iteration). Figure 4 shows an example of a 2-dimensional cellular automation. It has the dimension 8 x 5. Here, black cells are in state “1” whereas white cells are in state “0”.



Figures 4: Cellular Automation of Size 8 x 5

The new state of a cell is calculated based on its own current state and the current states of its neighbours. Figure 5 shows two of the most commonly used neighbourhoods for 2-dimensional cellular automata (Gerhardt and Schuster 1995). These neighbourhoods are the von-Neumann neighbourhood (Figure 5a) and the Moore neighbourhood (Figure 5b). Here, the grey cells belong to the neighbourhood of the black cell in the middle, i.e. they determine the next state of the cell in question.



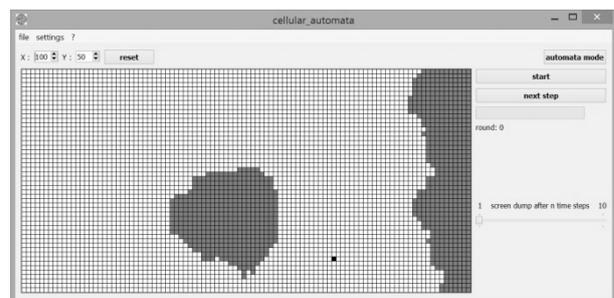
Figures 5: Common Neighbourhood Definitions for 2D Cellular Automata

The updating is performed following a set of application specific rules, i.e. these rules define the dynamic behaviour of the model.

Although cellular automata date back to the late 1940 (von Neumann 1966), they became popular after Conway presented his Game of Life in 1968 (Berlekamp et al. 1982). Conway developed this game to study complexity in nature whereas another well-known example of CA, the lattice gas cellular automation, was developed to simulate the physical world. Lattice gas CA was first introduced in 1973 by Hardy, Pomeau and de Pazzis. Their approach is known as HPP gas and is used to simulate fluid flows (Hardy et al. 1973). For a more comprehensive survey of the theory of cellular automata see for example (Kari 2005).

Simulation Tool Developed

The simulation tool that was developed in this work offers a variable size of the simulated environment (grid) and a graphical user interface (Figure 6).



Figures 6: Cellular Automata Tool

The software is written in C++ using Qt (Thein 2007) for the graphical user interface.

In the simulator, the user can set any particular cell to either *land*, *water* or *DOM source* simply by clicking the cell with the mouse. A scenario created in this way can then be saved for later use. The user can also define an interval after which the state of the world is saved to disk. This data can be used to visualize the evolution of the world over time. Last but not least, the world state can be saved as a C-function to be included in other C or C++ programs.

EXPERIMENTS

For the experiments, three different scenarios were simulated. The first one involves a coastline, one obstacle (island) and one DOM source (Figure 7). The second scenario is a variation of the first scenario with one additional obstacle (Figure 10) whereas the last scenario also includes an additional DOM source (Figure 13). The rules used in the simulations and the results of the experiments are presented below.

Rules

The cellular automation uses the Moore neighbourhood as described above. Each cell can be in one of 1 billion states, each represents one DOM concentration level L respectively the *land* state or the *DOM source* state.

The updating process is non-conventional in a way that a cell i is not updated by its neighbours. Instead, a cell updates all the cells in its neighbourhood; if the state of a neighbour is not the *land* state, the cell transfers part of its state (DOM level) ΔL_i to it (Equation 1). Each cell accumulates the donations it receives from all of its neighbours in order to determine its state for the next time step (Equation 2). Cells at the border treat the non-existing neighbours as sinks. As a consequence, state transferred to them is lost in the next iteration.

$$\Delta L_i(t) = \frac{L_i(t)}{n_i + 1} \quad (1)$$

Where:

$\Delta L_i(t)$: fraction of state at time step t
 $L_i(t)$: state of cell i at time step t
 n_i : neighbours of cell i that are not land or DOM source

$$L_i(t+1) = \Delta L_i(t) + \sum_{j=1}^{n_i} \Delta L_j(t) \quad (2)$$

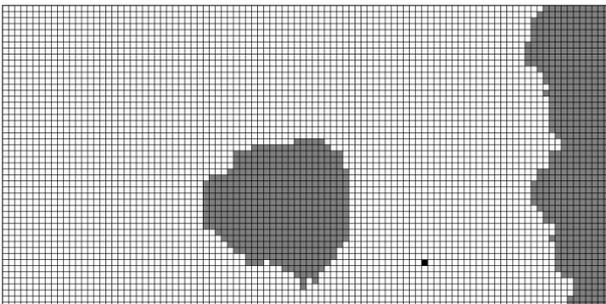
Where:

$L_i(t+1)$: state of cell i at time step $t+1$

The next section presents the results obtained by running three different scenarios.

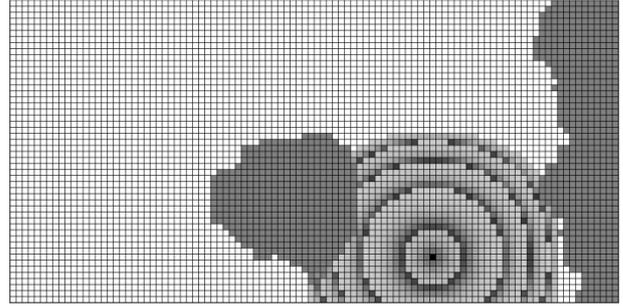
Results

Figure 7 shows the setup of the first scenario. It comprises of a coastline on the right hand side of the grid and an island in the lower middle part of the grid. There is also a single DOM source releasing a constant flow of DOM into the simulated environment.

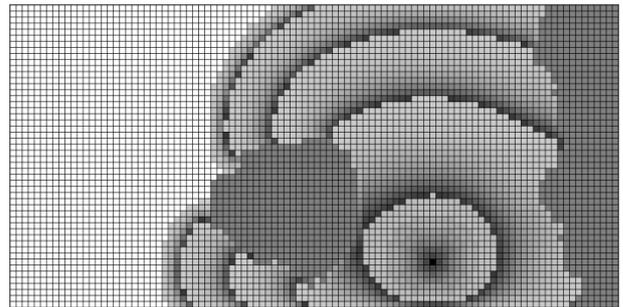


Figures 7: First Scenario at Time $t=0$

In Figures 8 and 9 show the concentration levels of DOM after time step 25 respectively 100.



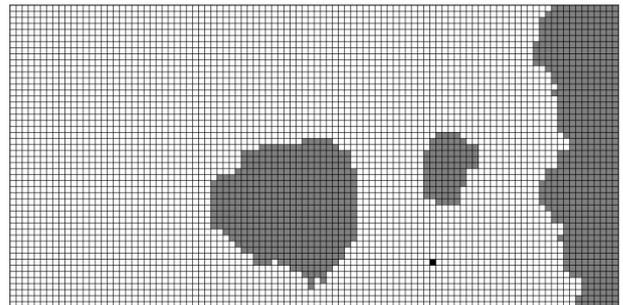
Figures 8: First Scenario at Time $t=25$



Figures 9: First Scenario at Time $t=100$

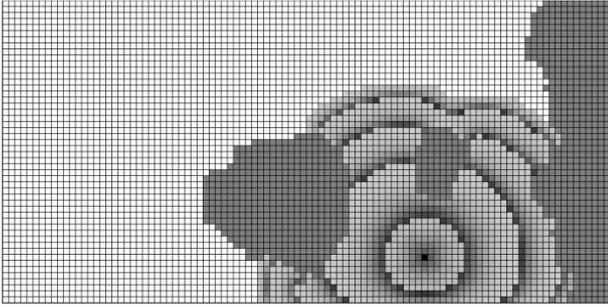
It can be seen that the DOM is diffused in the expected way. It eventually surrounds the island from both sides until it reaches a steady-state.

In the second scenario (Figure 10) a second obstacle (island) was introduced and the simulation was repeated.

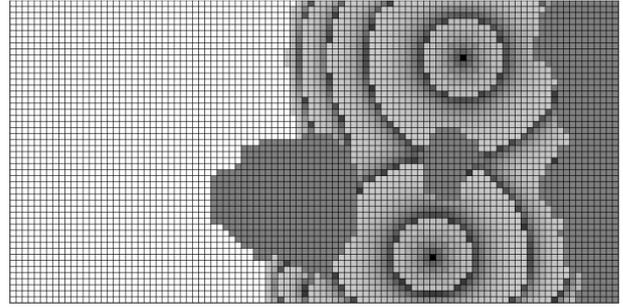


Figures 10: Second Scenario at Time $t=0$

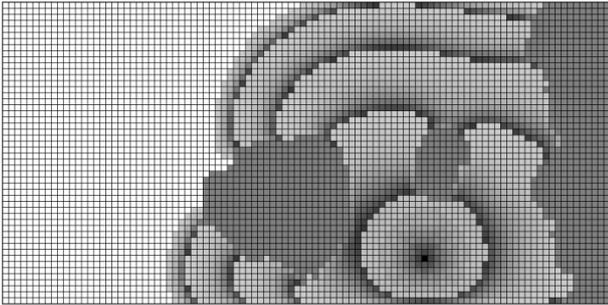
It can be seen from Figures 11 and 12 that the simulated DOM behaved also as expected: the DOM surrounds both islands and the shape of the concentration level is rather complex.



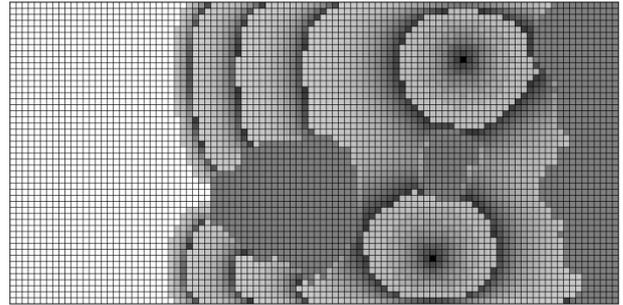
Figures 11: Second Scenario at Time $t=25$



Figures 14: Third Scenario at Time $t=25$

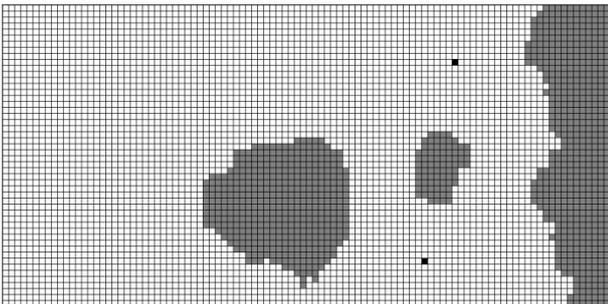


Figures 12: Second Scenario at Time $t=100$



Figures 15: Third Scenario at Time $t=100$

The third scenario differs from the second one in way that a second DOM source is added (Figure 13).



Figures 13: Third Scenario at Time $t=0$

Figures 14 and 15 show that here too the DOM concentration levels develop as expected.

CONCLUSIONS AND FUTURE WORK

The aim of this research was to develop a simulation program that can be used for the evaluation of search strategies for autonomous underwater vehicles. The simulator developed here is based on cellular automata. That means that both the search space and time are discretised. Also, it is limited to a 2-dimensional representation of the costal section to be simulated.

Results obtained from first experiments that were conducted using the simulator were promising and shoed expected behaviour. However, these results are only preliminary and qualitative. In order to improve the simulation, the rules need to be fine-tuned and the accuracy of the simulation needs to be analysed quantitatively.

The next step will be to include perturbation and wave mechanics to the model. The long term goal is to extend the simulator so that it works in three dimensions.

REFERENCES

- Berlekamp, E.R., Conway, J.H., Guy, R.K. 1982 *Winning Ways for Your Mathematical Plays II*, Academic Press.
- Gerhardt, M. Schuster, H. 1995 *Das digitale Universum*, Vieweg & Sohn.
- Hardy, Pomeau, J.Y., de Pazzis, O. 1973 "Time evolution of a two-dimensional model system. I. Invariant states and time correlation functions ", *Journal of Mathematical Physics*, Vol. 14, Issue 12, pp 1746-1759.
- Kari, J. 2005 "Theory of cellular automata:A survey", *Theoretical Computer Science*, Vol. 334, pp 3-33.
- Nolle, L. 2015 "On a search strategy for collaborating autonomous underwater vehicles", Proceedings of Mendel 2015, 21st International Conference on Soft Computing, Brno, CZ, pp 159-164.
- Stedmon, C.A., Markager, S., Rasmus Bro, B. 2003 "Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy", *Marine Chemistry*, Vol. 82, pp 239-254.
- Suryaputra, I.G.N.A. 2012 *Fluorescent dissolved organic matter (FDOM) in the coastal ocean: characterization, biogeochemical processes, and the possibility of in situ monitoring*, Doctoral Thesis, Carl von Ossietzky University, Oldenburg, Germany.
- Thelin, J. 2007 *Foundations of Qt Development*, Apress.
- von Neumann, J. in: Burks, A.W. (Ed.) 1966 *Theory of Self-Producing Automata*, University of Illinois Press.
- Wolfram, S. 1984 "Universality and complexity in cellular automata", *Physica D: Nonlinear Phenomena*, Vol. 10, Issue 1-2, pp 1-35.

AUTHOR BIOGRAPHIES

LARS NOLLE graduated from the University of Applied Science and Arts in Hanover, Germany, with a degree in Computer Science and Electronics. He obtained a PgD in Software and Systems Security and an MSc in Software Engineering from the University of Oxford as well as an MSc in Computing and a PhD in Applied Computational Intelligence from The Open University. He worked in the software industry before joining The Open University as a Research Fellow. He later became a Senior Lecturer in Computing at Nottingham Trent University and is now a Professor of Applied Computer Science at Jade University of Applied Sciences. His main research interests are computational optimisation methods for real-world scientific and engineering applications.

HOLGER THORMÄHLEN graduated from Jade University of Applied Science with a degree in Electrical Engineering and is currently studying for an MSc.

HARALD MUSA graduated from Fachhochschule Wilhelmshaven with a degree in Electrical Engineering. He subsequently obtained an MSc in Electrical Engineering from Jade University of Applied Science and is currently working as a Lecturer in the Department of Engineering Science.

DESIGN AND SIMULATION OF INTEGRATED EMI FILTER

Jens Werner*, Jennifer Schütt†, Guido Notermans†

*Jade University of Applied Science Wilhelmshaven/Oldenburg/Elsfleth, Friedrich-Paffrath-Str. 101, D-26389 Wilhelmshaven, Email: jens.werner@jade-hs.de

†NXP Semiconductors Germany GmbH, Stresemannallee 101, D-22529 Hamburg
Email: jennifer.schuett@nxp.com, guido.notermans@nxp.com

KEYWORDS

EM simulation, planar coils, method of moments, lumped model, common mode filter, USB 2.0, ESD protection, EMI

ABSTRACT

The design and simulation of an integrated common mode filter (CMF) for differential data lines, like the USB 2.0 interface, is presented in this paper. The device is manufactured in a bipolar semiconductor process for the integration of diodes protecting sensitive CMOS circuits against damage by Electrostatic Discharge (ESD). The filter is formed by planar coupled coils that are processed in copper/polyimide layers applied on top of silicon die. The design process is using 2.5D simulation techniques based on method of moments. Furthermore a lumped model is derived that allows exact and efficient transient simulations in SPICE [1] based simulators. The filter design itself shows strong common mode rejection in the GSM spectrum. The small size (1.34 mm x 0.95 mm) of the device makes it well suited for integration in modern mobile phone applications, to suppress electromagnetic interference (EMI) between a USB transmitter and a GSM receiver. Measurement data demonstrates the EMI protection in the GSM downlink spectrum. The dynamic resistance of the ESD diodes is derived in transmission line puls (TLP) measurements.

INTRODUCTION

Even though the specification of the USB 3.0 standard has been released in 2008, the adoption rate in the smartphone world is very low and USB 2.0 is still the dominant wire based interface for data exchange and for charging of mobile phones. With a natural data rate of 480 Mbit/s in high-speed mode, a certain spectral power density can be detected around 960 MHz due to harmonic content of the digital signals. This is depicted in Fig. 1: Multiple spurious signals close by the larger spur at 960 MHz appear as common mode component on a USB cable. The data on USB lines is NRZI (non-return-to-zero) encoded and bit stuffing is applied to ensure that data and clock are synchronous at the USB receiver. This allows data transmission without separate clock lines. The nature of the encoded data stream creates additional spectral components

below 960 MHz. In the standard GSM-900 system, 124 downlink channels are located between 935 MHz and 960 MHz. The extended system (E-GSM) adds additional channels at the lower end between 925 MHz and 935 MHz. The received signal strength can be close to the minimum sensitivity level of -102 dBm [2], e.g. in a rural area (with large distance between the transmitting base station and a mobile phone) or inside a building with shielding walls. Any additional noise from the internal USB interface that falls into the 200 kHz spectrum of a specific GSM downlink channel, will directly degrade the sensitivity. Since the GSM antenna (as potential EMI victim) is operating in an asymmetric mode and the high-speed data lines (as potential EMI source) are driven differentially, there is an inherent isolation. In real phone applications, the differential behaviour suffers partly from certain layout compromises (unequal line length, parasitic capacitances, etc.), partly from timing imperfections of the two single ended output driver circuits for D+ and D- signal lines (e.g. due to spread in semiconductor manufacturing). This results in a conversion from differential mode (DM) to common mode (CM) [3], [4] and provides the coupling path from USB to GSM antenna. For certain phone designs, this might already occur inside the phone (e.g. with long flexfoil cable between system on chip (SoC) and USB connector) or externally when a USB cable is plugged into the phone (see also Fig. 15). In order to increase the attenuation on this EMI path, a common mode filter can be placed on the USB data lines. The presented filter includes ESD protection diodes, which provide additional protection against damage from electrostatic discharge on the data lines, including the ID pin of USB on-the-go (OTG) ports [5].

In the development and manufacturing process of mobile phones many restrictions influence the physical shape, geometry and placement of components (e.g. main printed circuit board (PCB), battery, antennas, connectors and internal flexfoil interconnections). Those constraints are driven by aspects like size and cost reduction, EMI performance, antenna characteristics and usability. Fig. 2 shows three exemplary configurations, which differ in PCB geometry and placement of antenna module and USB port. The designs in Fig. 2 a) and b) allow for a very short routing between SoC and USB port and furthermore a maximized decoupling from the opposing

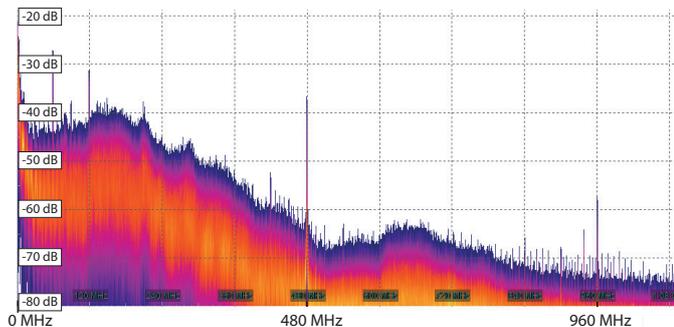


Fig. 1. Exemplary common mode (CM) spectrum on USB 2.0 data lines (measured during persistent reading from USB memory in high-speed mode, FFT-based spectrum by digital oscilloscope).

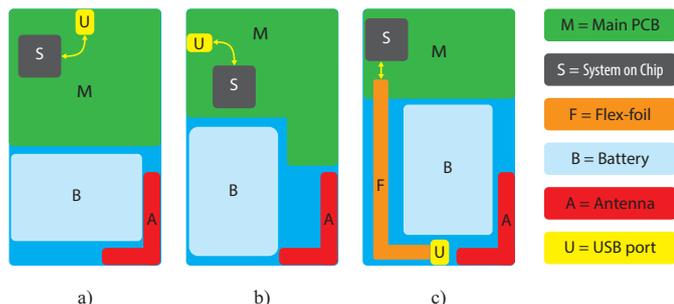


Fig. 2. Typical design variants found in mobile phones.

GSM antenna module. However, this position of the USB port might be inconvenient for the user: E.g. during the use of a car mount or docking station for battery charging and data synchronization via USB the placement at the bottom side is favorable (Fig. 2 c). An actual phone design that is using this architecture is depicted in Fig. 3. The long signal tracks of the flex-foil interconnection provide an internal coupling path into the GSM antenna structures. The close proximity of the USB port itself can exacerbate the coupling in particular when the quality of external cable plugs and the internal port has been deteriorated due to numerous insertion/removal cycles. In order to increase the isolation on this coupling path, a common mode filter can be used.

DESIGN AND SIMULATION

Circuit

The filter device is composed of two coupled planar coils and additional diode structures. Fig. 4 depicts the simplified schematic diagram, showing the fundamental elements and pin names of the USB connector. The design work flow is using electromagnetic (EM) simulation tools, based on the method of moments [6], [7], in order to consider the effects on the desired frequency response by distributed and parasitic structures. The ESD protection for the data lines is provided by diodes arranged in rail clamp structure; this reduces the capacitive load for the high-speed signals. Typical junction capacitance of reverse operated diode clamp is 1.5 pF. The



Fig. 3. Example of actual phone design with flexfoil USB connection.

concept of semiconductor controlled rectifiers (SCR) is applied to achieve a very low dynamic resistance [8], [9].

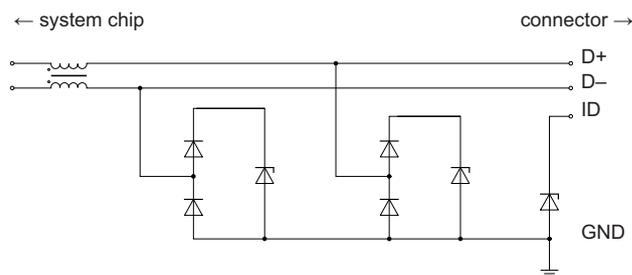


Fig. 4. Schematic diagram of common mode filter with ESD protection.

The coils are characterized by their series inductance $L_s \approx 60$ nH and series resistance $R_s \approx 6$ Ω . The coupling factor $k \approx 0.85$ provides sufficient DM passband to handle the spectrum of USB 2.0 signals. Inductance, coupling factor and diode capacitance shape the CM frequency response, that is designed to yield strong suppression in the 900 MHz GSM band. Nevertheless, it does not affect the single ended transmission as used for low-speed and full-speed signaling (USB 1.0/1.1).

Once the design was finalized, a more detailed lumped element equivalent circuit has been created for the purpose of signal integrity studies. This passive circuit model is rather compact, yet allows high accuracy for efficient transient simulations like eye diagrams. Compared with S-parameters the lumped model can reflect the device behaviour exactly at DC and low frequencies. The basic cell is shown in Fig. 5. In combination with few additional components, it is used four times to simulate the distributed characteristic of the real device for frequencies from 0 Hz beyond 2 GHz.

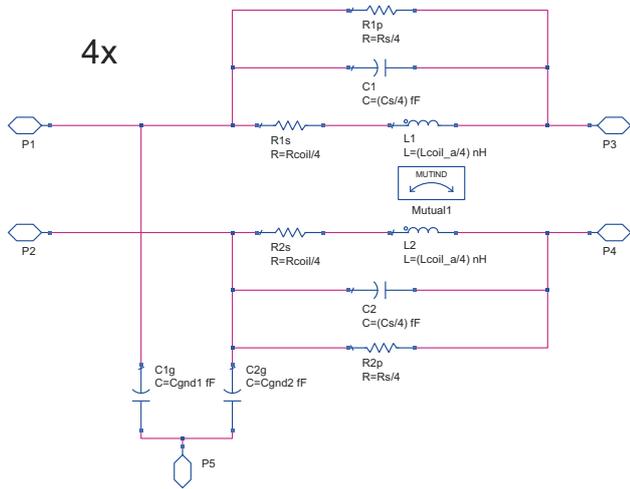


Fig. 5. Schematic diagram of basic element for lumped model.

Implementation

The ESD diode structures are manufactured in a bipolar semiconductor process with two aluminium layers. The die size is 1.34 mm x 0.95 mm. Layers of polyimid and copper are processed on top of the silicon die in order to implement the two coupled coils with acceptable series resistance. In a final processing step solder balls are added. In Fig. 6 these solder balls face up, and the upper copper layer is clearly recognizable. The lower copper layer is hardly distinguishable in this photograph. The cross sectional views in Fig. 7 and Fig. 8 provide a better view on the two layers. The former is taken from the EM simulation tool [6] and reveals a cross-section, that allows to identify both copper layers and the electrical connection between the coils and pads for the solder balls. The latter was taken by a scanning electron microscope (SEM). The transversal dimension of a single coil winding is about 8 μm by 8 μm .

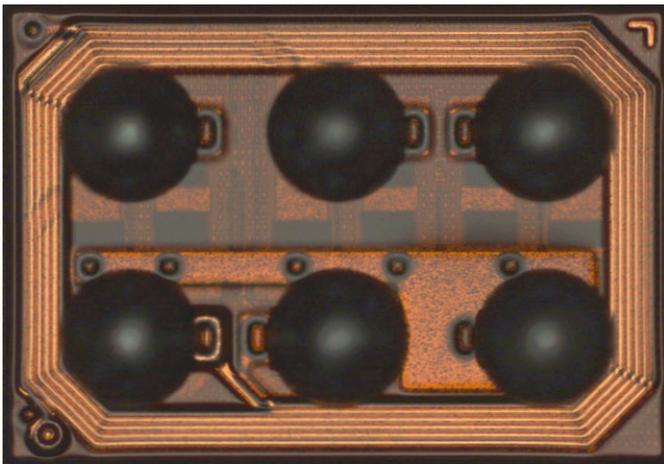


Fig. 6. Photograph of silicon die, showing solder balls, copper layers and underlying aluminium layer on top of semiconductor substrate.

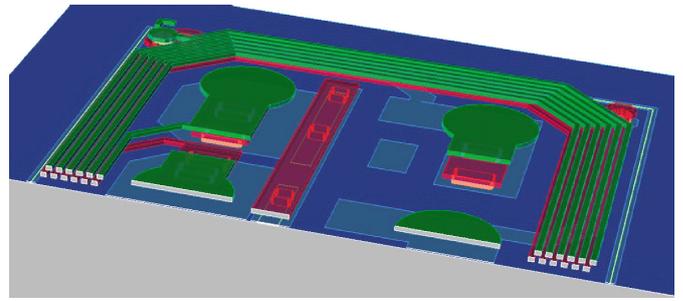


Fig. 7. Cross sectional view from EM simulation tool: Copper layers in red/green, aluminium layer in light blue, solder balls are not visible.

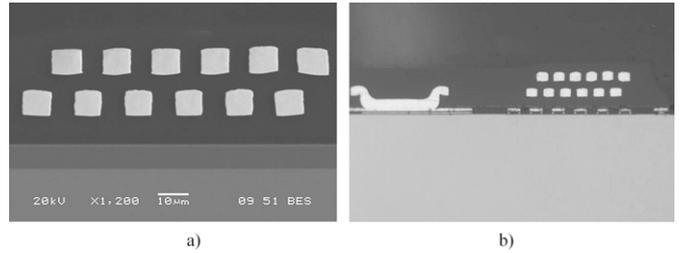


Fig. 8. Cross sectional view from SEM: Upper and lower copper coil layer (a), larger view showing interconnect from aluminium lower copper layer (b)

RESULTS

ESD protection

The ideal ESD protection is a switch with an infinite on-resistance in normal operation and zero on-resistance during an ESD event. A semiconductor controlled rectifier (SCR) is near-ideal in the sense that it exhibits a very high resistance for low positive voltages and switches to a highly-conductive state after triggering. Fig. 9 a) shows a typical SCR I-V graph. Once a certain voltage (trigger voltage V_{t1}) is exceeded, the SCR triggers and the voltage drops to a very low value, the holding voltage V_h , which is typically 1.5 V ... 2 V. During the transition from V_{t1} to V_h , the SCR exhibits briefly a negative dynamic resistance. An equivalent circuit diagram is shown in Fig. 9 b). Beyond V_h , the voltage rises again with increasing

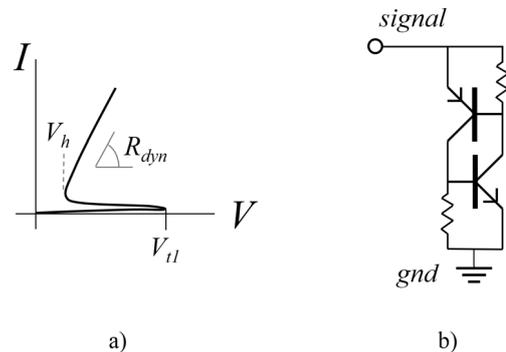


Fig. 9. SCR I-V graph (left); SCR equivalent circuit diagram (right).

current in accordance with the ohmic resistance R_{dyn} of the device. An SCR typically has a very low dynamic resistance

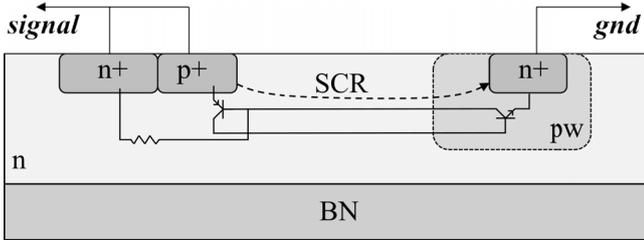


Fig. 10. Cross-section of an SCR showing the integrated return diode.

in this range, in the order of a few tenths of an Ohm. Thus, the power dissipation in an SCR is very low during an ESD event, resulting in very high current capability in case of an ESD pulse, up to 80 mA/ μm width of the SCR.

For reverse polarity ESD pulses, the ESD current is handled by the diode from D+ respectively D- to GND (Fig. 4). The overall concept is a Low-Voltage-Triggered SCR (LVTSCR), [10], [11], in which the low-voltage triggering is not achieved by means of an avalanche diode. A cross-section of the device is shown in Fig. 10. For positive signal voltages the SCR is formed by the p+ in the n-epi, the pwell (pw) and finally the n+ to ground (gnd).

Both, the low clamping voltage and the low dynamic resistance of the SCR contribute to an excellent protection of the SoC in the following way:

- 1) The low clamping voltage reduces the risk of over-voltage on the SoC input/output (IO) pins, which might otherwise damage the sensitive gate oxides, which in advanced CMOS devices can be destroyed at voltages below 5 V.
- 2) The low dynamic resistance shunts most of the ESD current to ground. The residual current into the SoC decreases, as the SCR dynamic resistance gets lower. By a proper choice of the protection, the SoC current capability and the impedance in between, the residual current into the SoC can be limited to values preventing thermal damage to the SoC [12].

The current capability of all components can be quantified according to a number of international standards (e.g. HBM [13]). The system ESD performance can be established using an IEC 61000-4-2 test [14]. Such tests yield a pass/fail result, which is useful for production-type testing but does not give much support for the development process. During development, the Transmission Line Pulse (TLP [15]–[17]) is widely used. A TLP system applies square current pulses, usually of 100 ns duration, to the device under test (DUT) which allow accurate determination of the voltage and current during the pulse. The TLP pulses provide an ESD-like discharge (regarding the pulse duration) and are known to correlate very well with HBM and IEC pulses [18], [19]. Fig. 11 shows a typical TLP pulse of about 0.4 A of current and the resulting voltage pulse. The voltage pulse shows, that when a current pulse is applied, the voltage briefly (< 0.5 ns) rises to the trigger voltage V_{t1} and then drops sharply to the holding voltage V_h .

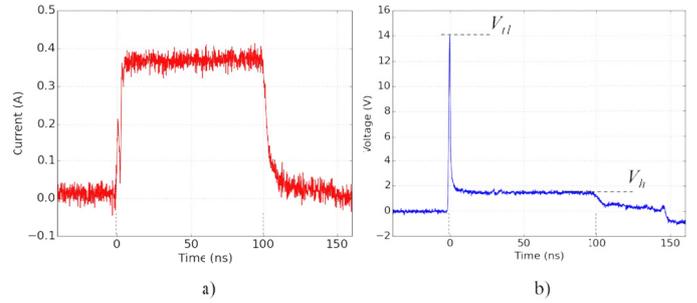


Fig. 11. Typical 100 ns TLP current (left) and voltage pulse (right).

Because the TLP employs pulses in the ESD time frame, it is possible to measure the I-V curve under typical ESD conditions, regarding pulse width (100 ns) and rise time (10 ns). A full TLP I-V curve is measured by applying a succession of pulses with increasing current until a pre-determined maximum current is reached or the device fails. The time interval between two consecutive pulses is typically about 1 s, which is much larger than the cool-down time of the device (which is in the order of ns). Thus, cumulative heating effects are avoided. As shown in Fig. 11 a), each current pulse exhibits a flat plateau of about 100 ns after the initial transient during the rise time of the pulse. An HPPI 3010C TLP system - combined with a 3011C pulse width extender - was used to apply the TLP pulses. Voltage and current within each pulse were measured on-wafer by means of RF Kelvin probes connected to a 4 GHz Tektronix TDS7404B digital oscilloscope (at 10 GS/s). Each current and voltage measurement represents the average of the signal at the plateau between 75-85 ns during the pulse width. The force signal is applied into 50 Ohms to the DUT and the voltage sense signal is picked up via a 5 kOhm series resistor. The current is measured by means of a Tektronix CT1 current probe, connected via an attenuator of 26 dB. By successively increasing the amplitude of each current pulse, a complete I-V curve can be measured. Failure of the device can be detected by monitoring the DC leakage current at +5 V between two consecutive pulses, using a Keithley 2636 SMU. Fig. 12 shows an example of a TLP I-V curve for an SCR as implemented in this common mode filter. The curve appears continuous, because the data points are connected by a solid line and the individual pulse measurements are not indicated, but in fact it represents a succession of individual pulses. The leakage measurements are not shown, because the device did not fail within the pulse current range of -12 A to +12 A as depicted in Fig. 12. Using this I-V curve, it is straightforward to measure device parameters, such as the dynamic resistance R_{dyn} . Such a measurement cannot be performed using DC measurements, because the much longer pulse durations would lead to extreme power dissipation and thus to premature failure unrelated to the actual ESD performance. The TLP measurement results in Fig. 12 cover both polarities: For pulses with negative polarity the ground diode gets conductive once the voltage exceeds the forward voltage ($V_{cl} < -1$ V). The dynamic resistance during the TLP

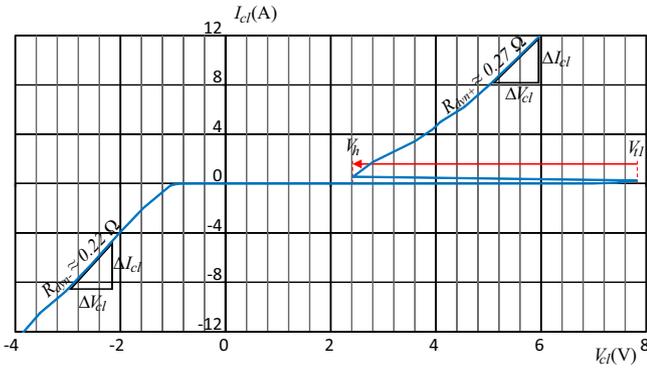


Fig. 12. I-V graph of the integrated SCR, measured with 100 ns TLP.

pulse can be read from the slope in Fig. 12 as $R_{dyn-} \approx 0.22\Omega$. For positive polarity the SCR is triggered at $V_{t1} = V_{cl} \approx 8$ V. After this snap-back, the SCR provides a dynamic resistance of $R_{dyn+} \approx 0.27\Omega$.

Mixed mode scattering parameters

Scattering parameters are measured using a four port vector network analyzer (VNA) and converted into mixed mode S-parameters S_{dd21} (solid lines) and S_{cc21} (dashed lines) as shown in Fig. 13 [20], [21]. Besides the measured data (blue curves), the graph includes simulation results based on the EM model (red curves) and the lumped model (green curves).

For low frequencies all six curves show an insertion loss of $S_{21} \approx -0.5$ dB due to the series resistance of the coils. The -3 dB cut-off frequency for the differential mode is above 1 GHz, for the common mode below 200 MHz. The resonant behaviour of inductive and capacitive elements in the filter design results in a significant common mode rejection over the full GSM-900 uplink (880 MHz to 915 MHz) and downlink bands (925 MHz to 960 MHz).

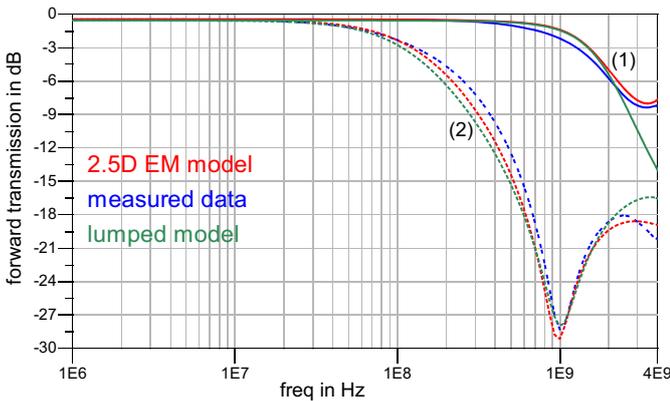


Fig. 13. Measured mixed-mode S-parameter:(1) = S_{dd21} ; (2) = S_{cc21} .

Signal integrity

The eye diagram is used to assess the signal integrity in the time domain. The USB specification defines various limits for the eye diagram to be measured at different test points in a

system. The most severe, i.e. the widest eye opening, is given by template 1 as specified in [22]. The eye diagram as depicted in Fig. 14 was measured with the filter device mounted in a complete application circuit, proving that insertion loss and cut-off frequency in differential mode are compliant with USB 2.0 requirements.

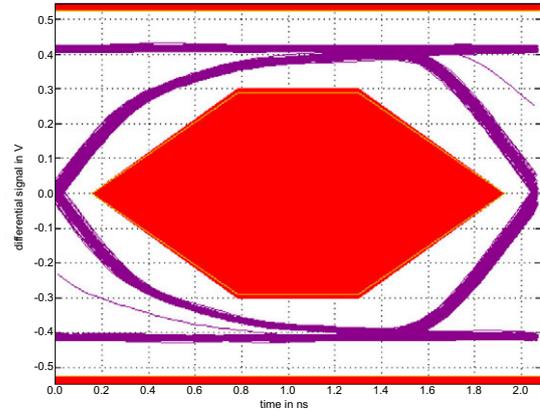


Fig. 14. Measured eye diagram according to template 1, test point TP2 [22].

Impact on USB EMI emission

As discussed above, several coupling paths from the USB interface to the GSM antenna module might exist in a mobile phone (Fig. 15). Investigations based on commercially available phones can be difficult as documentation of hardware and software is hardly available. Thus, a Beagleboard running the Linux operating system is used. The SoC incorporates an AM335x Cortex A8 ARM processor. This SoC is similar to those used in modern smartphones and provides two integrated USB 2.0 OTG host controllers. One of them is connected to a female type A USB connector, which is used during testing.

1) **Test setup:** Since the application board is operated with opensource software, full access to the low level USB functions is available. This is essential for good reproducibility of persistent USB transmissions. Here “persistent” means that data is read continuously from an external USB memory as fast as possible, without additional delay from higher software layers, circumventing any caching that would reduce or avoid the physical USB transmission. The data itself is still transmitted in packets. This behaviour creates a non-stationary frequency spectrum. In order to observe this spectrum with a swept spectrum analyzer, the maximum hold detector was used over ten consecutive sweeps.

To mimic an arbitrary coupling path between the USB data lines and an asymmetric (antenna) structure a single winding current probe is used: The inner conductor of a coaxial cable is twined around the paired data lines and soldered to the outer conductor [23]. In a real phone, this coaxial cable would be connected to the RF frontend of the GSM module; here, an EMI test receiver measures the spectrum in the GSM downlink channels (compare Fig. 15 and 16). In this experiment, only the single coupling path between an external USB cable and the asymmetric probe is evaluated (Fig. 16).

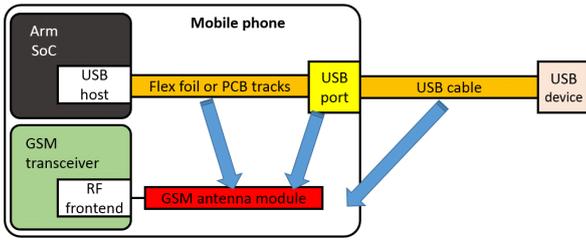


Fig. 15. Potential EMI interference paths from USB interface to GSM antenna module in a mobile phone.

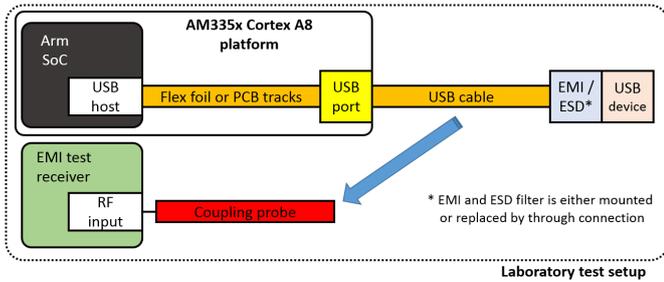


Fig. 16. Laboratory test setup to measure impact of EMI filter during persistent USB traffic and emulate condition in mobile phones.

Fig. 17 depicts the filter PCB with the common mode filter (detail 1) in blue frame) in front of a USB memory device. During testing, this PCB is replaced by an identical one, but without a filter device. Here the filter footprint is bypassed by wires (detail 2) in orange frame) in order to establish a through connection. Any impact of the filter PCB itself, such as reflections at the male and female ports, is therefore kept the same. Finally the spectrum in the GSM downlink band is measured with and without the filter device.

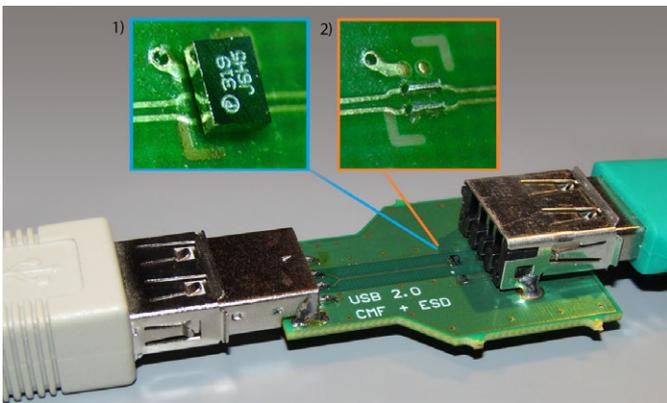


Fig. 17. The filter PCB is attached to a USB memory stick. Detail 1: CMF is mounted. Detail 2: CMF is replaced by through connection.

2) **Results:** The observed spectrum is given in Fig. 18: The orange graph is measured without and the blue graph with the common mode filter in place. Both graphs show a strong signal at 960 MHz (Marker A) which is linked to the harmonic spectrum of the inherent clock signal for 480 MBit/s USB data signals (see also Fig. 1). This noise

can be particularly detrimental to absolute radio-frequency channel numbers (ARFCN) 123 ($f_c = 959.6$ MHz) and 124 ($f_c = 959.8$ MHz).

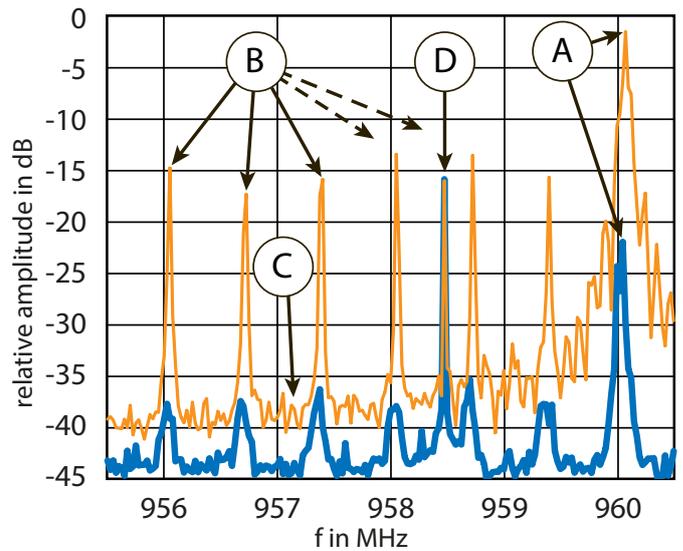


Fig. 18. Detailed view on GSM downlink spectrum measured with (lower, blue graph) and without (upper, orange graph) common mode filter during persistent read access on USB memory (maximum hold detector over 10 sweeps).

Furthermore multiple narrowband interferers (Marker B) can be identified, spaced by 0.67 MHz offset. Given the 200 kHz channel spacing in GSM, almost every third downlink channel would be impaired by these spurs. By adding the CMF into the USB transmission path, all of these spurs are reduced by approximately 20 dB. This corresponds well with the measured S-parameter S_{cc21} (Fig. 13). Furthermore this proves that the coupling path between USB cable and coupling probe is dominated by common mode. Likewise the noise spectrum (Marker C) is reduced, although the reduction by 20 dB is masked due to sensitivity limitations of the chosen test setup (impact of maximum hold detector and insertion loss of coupling probe). The spur denoted by marker D is not affected by the CMF. It is identified as non-USB related, direct EMI injection from the Beagleboard into the coupling probe: even without any USB transmission and disconnected USB memory this spurious emission can be measured.

CONCLUSION

The design process for a common mode filter with integrated ESD protection for the USB 2.0 interface is presented in this paper. The EM simulation, based on 2.5D tools, has demonstrated its strength in simulating planar structures in semiconductor devices. Simulated and measured mixed-mode scattering parameters show very good conformance. The achieved frequency response allows to suppress common mode signals in the GSM-900 downlink spectrum as demonstrated by an experimental test setup. The ESD protection is realized by advanced SCR structures that protect sensitive CMOS SoCs due to their low clamping voltage and low dynamic resistance.

REFERENCES

- [1] L. W. Nagel, "What's in a name?" *IEEE Solid-State Circuits Magazine*, vol. 3, no. 2, pp. 8–13, Spring 2011.
- [2] 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Radio transmission and reception, "3GPP TS 05.05 V8.20.0," 1999.
- [3] F. De Paulis, L. Raimondo, D. Di Febo, and A. Orlandi, "Routing Strategies for Improving Common Mode Filter Performances in High Speed Digital Differential Interconnects," in *Signal Propagation on Interconnects (SPI)*, 2011 15th IEEE Workshop on, May 2011, pp. 3–6.
- [4] Yu-Jen Cheng, Hao-Hsiang Chuang, Chung-Kuan Cheng, and Tzong-Lin Wu, "Novel Differential-Mode Equalizer With Broadband Common-Mode Filtering for Gb/s Differential-Signal Transmission," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 9, pp. 1578–1587, 2013.
- [5] "On-The-Go and Embedded Host Supplement to the USB 2.0 Specification, Revision 2.0," May 2009.
- [6] Keysight. (2015) ADS - Advanced Design System. [2016-03-14]. [Online]. Available: <http://www.keysight.com/find/eesof-ads>
- [7] R. F. Harrington, *Field Computation by Moment Methods*. The Macmillan Company, 1968.
- [8] A. Amerasekera and C. Duvvury, *ESD in Silicon Integrated Circuits*, 2nd ed. New York, John Wiley & Sons, 2002.
- [9] J. Di Sarro and E. Rosenbaum, "A Scalable SCR Compact Model for ESD Circuit Simulation," *Electron Devices, IEEE Transactions on*, vol. 57, no. 12, pp. 3275–3286, Dec 2010.
- [10] A. Chatterjee and T. Polgreen, "A low voltage triggering SCR for on-chip ESD protection at output and input pads," *IEEE Elec. Dev. Lett.*, vol. 21, no. 12, 1991.
- [11] G. Notermans, F. Kuper, and J.-M. Luchies, "Using an SCR as ESD protection without latch-up danger," vol. 37, no. 10, pp. 1457–1460, 1997.
- [12] "System level ESD: Part II: Implementation of effective ESD robust designs." JEDEC Publication JEP162, 2013.
- [13] "Electrostatic Discharge Sensitivity Testing - Human Body Model (HBM) - Component Level." ESDA/JEDEC JS-001, 2014.
- [14] "Electromagnetic Compatibility (EMC) - Part 4-2: Testing and measurement techniques - Electrostatic discharge immunity test." IEC 61000-4-2, 2008.
- [15] T. J. Maloney and N. Khurana, "Transmission Line Pulsing Techniques for Circuit Modeling of ESD Phenomena," in *EOS/ESD Symposium Proceedings*, 1985, pp. 49–54.
- [16] H. Hyatt, J. Harris, A. Alonzo, and P. Bellew, "TLP Measurements for Verification of ESD Protection Device Response," *Electronics Packaging Manufacturing, IEEE Transactions on*, vol. 24, no. 2, pp. 90–98, Apr 2001.
- [17] S. Voldman, R. Ashton, J. Barth, D. Bennett, J. Bernier, M. Chaine, J. Daughton, E. Grund, M. Farris, H. Gieser, L. Henry, M. Hopkins, H. Hyatt, M. Natarajan, P. Juliano, T. Maloney, B. McCaffrey, L. Ting, and E. Worley, "Standardization of the Transmission Line Pulse (TLP) Methodology for Electrostatic Discharge (ESD)," in *Electrical Overstress/Electrostatic Discharge Symposium, 2003. EOS/ESD '03.*, Sept 2003, pp. 1–10.
- [18] G. Notermans, P. D. Jong, and F. Kuper, "Pitfalls when correlating TLP, HBM and MM testing," in *Electrical Overstress/Electrostatic Discharge Symposium Proceedings, 1998*, Oct 1998, pp. 170–176.
- [19] G. Notermans, S. Bychikhin, D. Pogany, D. Johnsson, and D. Maksimovic, "HMMTLP correlation for system-efficient ESD design," in *Microelectronics Reliability*, vol. 52, 2012, pp. 1012–1019.
- [20] D. Bockelman and W. Eisenstadt, "Combined Differential and Common-Mode Scattering Parameters: Theory and Simulation," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 43, no. 7, pp. 1530–1539, Jul 1995.
- [21] M. K. Allan Huynh and S. Gong. (2010) Mixed-Mode S-Parameters and Conversion Techniques, *Advanced Microwave Circuits and Systems*. [2015-05-28]. [Online]. Available: <http://www.intechopen.com/books/advanced-microwave-circuits-and-systems/mixed-mode-s-parameters-and-conversion-techniques>
- [22] "Universal serial bus specification, Revision 2.0," Apr. 2000.
- [23] D. Morgan, *A Handbook for EMC Testing and Measurement (IET Electrical Measurement Series)*, 1st ed. Institution of Engineering and Technology, 1994.



Jens Werner was born in Cologne, Germany in 1969. He received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the Technical University of Braunschweig, Braunschweig, in 1996 and 2002, respectively. In 1996 he was working with Aerodata AG as a Flight Inspection Engineer working on calibration of airborne antennas. From 1996 to 2001, he was a Research Assistant at the Institute of Electromagnetic Compatibility, Technical University of Braunschweig. His main research interests were measurement techniques and representation of guided and radiated electromagnetic fields. In 2001 he joined the Innovation Centre of Philips Semiconductors Germany GmbH in Hamburg, (since 2006 NXP Semiconductors). In March 2014, he became a Professor at Jade University of Applied Sciences, Wilhelmshaven, Germany. He is responsible, amongst others, for the laboratory for RF, Wireless and EMC.



Jennifer Schütt was born in Heide, Germany in 1981. She received a degree (Dipl.-Ing.) in electrical engineering from Technical University of Hamburg (TUHH) in 2009. Since then she is with NXP Semiconductors, Hamburg, developing EMI-Filter and ESD protection devices. Main field of expertise is in device modelling, EM simulation, device physics and project management. Currently she is working on common-mode-filter designs with integrated ESD protection for ultra-fast differential data lines. In the field of EM simulation, she organizes regular cross team RF-expert meetings for NXP engineers located in Hamburg.



Guido Notermans is senior principal ESD with NXP Semiconductors in Hamburg. He graduated in experimental physics at Utrecht University in 1980 and received his PhD in plasma physics in 1984. He subsequently joined Philips Semiconductors where he developed III-V semiconductor lasers until 1990, first at the Philips Research Labs in Eindhoven. From 1995 he worked as senior ESD principal for Philips Semiconductors. In 1999 he moved to Berlin where he joined Infineon Fiber Optics as R&D manager for electro-optical devices. In 2005 he moved to Philips Semiconductors Zurich where he returned to the field of ESD. Philips Semiconductors was spun off to become NXP in 2006. In 2013, Guido moved to Hamburg and is developing stand-alone (off-chip) ESD protection in the BU Standard Products.

Modelling, Simulation and Control of Technological Processes

IDENTIFICATION AND LQ DIGITAL CONTROL OF A SET OF EQUAL CYLINDER ATMOSPHERIC TANKS – SIMULATION STUDY

Vladimír Bobál, Petr Dostál, Marek Kubalčík and Stanislav Talaš
Tomas Bata University in Zlín
Faculty of Applied Informatics
Nad Stráněmí 4511
760 05 Zlín
Czech Republic
E-mail: bobal@fai.utb.cz

KEYWORDS

High-order process, Time-delay system, Set of liquid tanks, Smith predictor, LQ digital control, Simulation of control loops.

ABSTRACT

Time-delays (dead time) are found in many processes in industrial practice. Time-delays are mainly caused by the time required to transport mass, energy and information. In many cases time-delay is caused by the effect produced by the accumulation of a large number of low-order systems. One of possibilities of control of such processes is their approximation by lower-order model with time-delay. The contribution is focused on the design of an algorithm for digital control of high-order process that is approximated by a second-order model with time-delay. The controller algorithms use the digital modification of the linear quadratic (LQ) Smith predictor (SP). The LQ criterion was combined with pole assignment principle. These algorithms were applied to the control of a set of equal liquid cylinder atmospheric tanks.

INTRODUCTION

Some technological processes in industry are characterized by high-order dynamic behaviour or large time constants and time-delays. For control engineering, such processes can often be approximated by the FOTD (first-order-time-delay) model. Time-delay in a process increases the difficulty of controlling it. However using the approximation of a high-order process by a lower-order model with time-delay provides simplification of the control algorithms. Let us consider a continuous-time dynamical linear SISO (single input $u(t)$ – single output $y(t)$) system with time-delay L . The transfer function of a pure transportation lag is e^{-Ls} where s is a complex variable. Overall transfer function with time-delay is in the form

$$G_L(s) = G(s)e^{-Ls} \quad (1)$$

where $G(s)$ is the transfer function without time-delay.

Processes with time-delay are difficult to control using standard feedback controllers. When a high performance of the control process is desired or the relative time-delay is very large, a predictive control strategy must be used. The predictive control strategy 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 contained a dynamic model of the time-delay process and it can be considered as the first model predictive algorithm.

Historically first modifications of time-delay algorithms were proposed for continuous-time (analogue) controllers using various approaches. In industrial practice, the implementation of the time-delay compensation algorithms on continuous technique is difficult.

One of possible approaches to control of process with time-delay is digital Smith predictor based on polynomial theory.

Polynomial methods are design techniques for complex systems (including multivariable), signals and processes encountered in control, communications and computing that are based on manipulations and equations with polynomials, polynomial matrices and similar objects. Systems are described by input-output relations in fractional form and processed using algebraic methodology and tools (Šebek and Hromčík 2007). Controller design consists in solving polynomial (Diophantine) equations. This paper is oriented to design of a robust LQ control using polynomial theory. The Diophantine equations can be solved using the uncertain coefficient method – which is based on comparing coefficients of the same power. This is transformed into a system of linear algebraic equations (Kučera 1993).

The digital pole assignment Smith predictor was designed using a polynomial approach in (Bobál et al. 2011). The design of this controller was extended by a method for a choice of a suitable pole assignment of the characteristic polynomial. Because the classical analogue Smith predictor is not suitable for control of unstable and integrating time-delay processes, the polynomial digital LQ Smith predictor for control of unstable and integrating time-delay processes has been designed in (Bobál et al. 2014).

It is obvious that the majority processes met in industrial practice are influenced by uncertainties. The uncertainties suppression can be solved either implementation adaptive control or robust control. Some adaptive (self-tuning) modifications of the digital Smith predictors are designed in (Hang et al. (1989; 1993); Bobál et al. 2011). Two versions of these controllers were implemented into MATLAB Toolbox (Bobál et al. 2012a; Bobál et al. 2012b).

The paper is organized in the following way. The general problem of a control of the time-delay systems with regard to polynomial approach is described in Section 1. The fundamental principle of digital Smith predictor is described in Section 2. The high-order system (a set of n equal liquid cylinder atmospheric tanks) is analysed in Section 3. Section 4 contains description of identification procedures. Two versions of the primary polynomial LQ controller, which are components of the digital Smith Predictor, are proposed in Section 5. The simulation verifications of individual control-loops with their results are presented in Section 6. Section 7 concludes this paper.

PRINCIPLE OF DIGITAL SMITH PREDICTOR

The discrete versions of the SP and its modifications are more suitable for time-delay compensation in industrial practice. The block diagram of a digital SP (see Hang et al. (1989, 1993)) is shown in Fig. 1. The function of the digital version is similar to the classical analogue version.

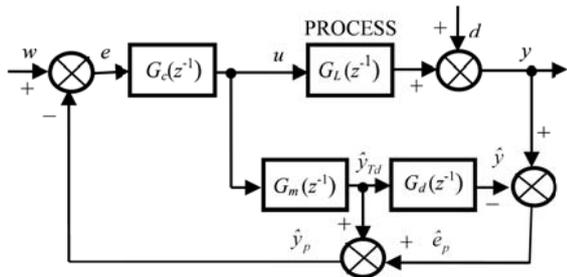


Figure 1: Block diagram of a digital Smith Predictor

Number of high-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

$$G_L(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} z^{-d} = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} z^{-d} \quad (1)$$

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. The block $G_m(z^{-1})$ represents process dynamics without the time-delay and is used to compute an open-loop prediction. The numerator in transfer function $G_d(z^{-1})$ is replaced by its static gain

$B(1)$, i.e. for $z = 1$. This is to avoid problem of controlling a model with a $B(z^{-1})$, which has non-minimum phase zeros caused by a high sampling period or fractional delay. Since $B(z^{-1})$ is not controllable as in the case of a time-delay, it is moved out of the prediction model $G_m(z^{-1})$ and is treated together with the time-delay block, as shown in Fig. 1. 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 d are the error and the measured disturbance, w is the reference signal. The primary (main) controller $G_c(z^{-1})$ can be designed by different approaches (for example digital PID control or methods based on polynomial approach). The detailed description of the block diagram (Fig. 1) is in (Bobál et al. 2011).

SERIES OF EQUAL LIQUID TANKS

In many cases in industrial practice the time-delay is caused by the effect produced by the accumulation of a large number of low-order systems. Consider a set of n equal cylinder atmospheric tanks, where a single tank is shown in Fig. 2 (Torrico and Normey-Rico 2007) and the whole set is shown in Fig. 3. In this system, the output flow of tank i (q_{iO}) feeds tank $i + 1$; that is, the input flow tank $i + 1$ is $q_{(i+1)I} = q_{iO}$. If all the tanks have the same area (F) of crosscut and the individual tank levels are near to an operating point, then the dynamic behaviour of the level in each tank h_i can be modelled by a linear system

$$\begin{aligned} F \frac{dh_i}{dt} &= q_{iI} - q_{iO} \\ q_{iO} &= K_1 h_i \end{aligned} \quad (2)$$

where K_1 is a constant that depends on the tank characteristics.

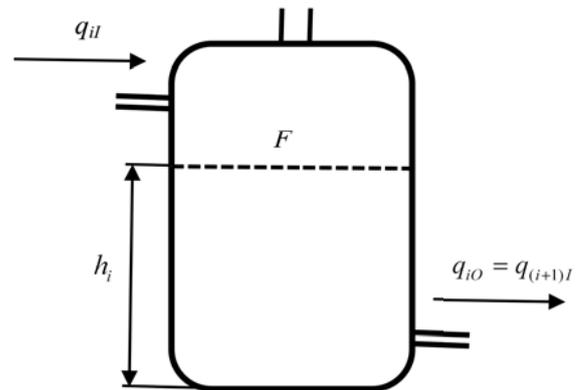


Figure 2: Schema of liquid cylinder tank

Consider a set of n tanks as shown in Fig. 3. Thus, the transfer function relating the input follow in tank i and its level is given by

$$h_i(s) = \frac{1/K_1}{Ts + 1} q_{iI}(s) \quad (3)$$

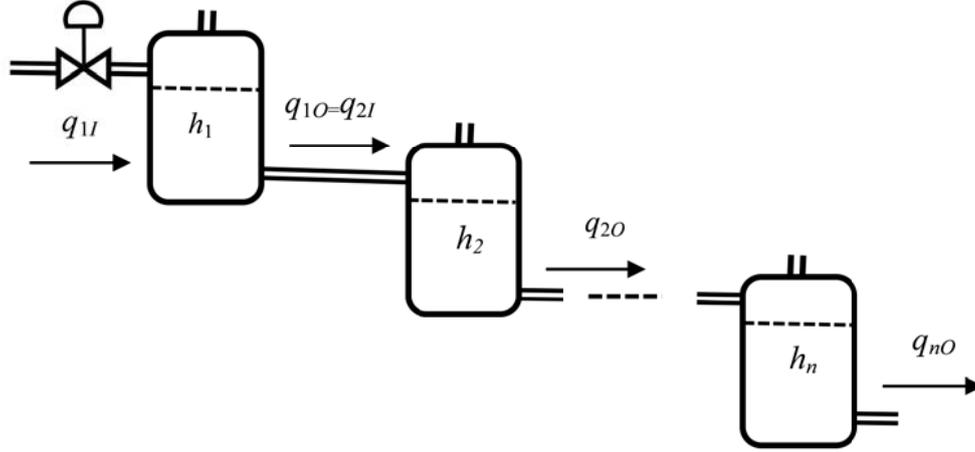


Figure 3: Series of liquid cylinder tanks

where $T = F / K_1$ is time constant.

For tank 1 is

$$h_1(s) = \frac{1/K_1}{Ts+1} q_{1I}(s) \quad (4)$$

and for tank 2 using the second equation of (2)

$$h_2(s) = \frac{1/K_1}{Ts+1} q_{2I}(s) = \frac{1/K_1}{Ts+1} q_{1O}(s) = \frac{1/K_1}{Ts+1} K_1 h_1(s) \quad (5)$$

Then, using the expression (4) it follows

$$h_n(s) = G(s) q_{1I}(s) = \frac{K_g}{(Ts+1)^n} q_{1I}(s) \quad (6)$$

and the transfer function of the series of tanks system is

$$G(s) = \frac{h_n(s)}{q_{1I}(s)} = \frac{K_g}{(Ts+1)^n} \quad (7)$$

where $K_g = 1/K_1$ is static gain of the system.

Consider for simulation experiments of control model (7) the eight – order system, i. e. $n = 8$. Following parameters of the individual liquid tanks are considered (see Fig. 1): high of tank $h = 1.5$ m; diameter of tank

$d_T = 1$ m; tank area $F = \frac{\pi d_T^2}{4} = 0.785 \text{ m}^2$; set point

$h_1 = 1$ m; time constant $T = 2$ min;

constant $K_1 = \frac{F}{T} = \frac{0.785}{2} = 0.3925 \text{ m}^2 \text{ min}^{-1}$;

static gain $K_g = \frac{1}{K_1} = \frac{1}{0.3925} = 3.08 \text{ m}^2 \text{ min}$.

The resulting transfer function is given by

$$G(s) = \frac{h_8(s)}{q_{1I}(s)} = \frac{3.08}{(2s+1)^8} \quad (8)$$

If (8) is the transfer function of a continuous-time dynamic system, then the following expression for the

discrete transfer function with zero-order holder and sampling period T_0 is valid

$$G(z^{-1}) = \frac{b_1 z^{-1} + b_2 z^{-2} + \dots + b_8 z^{-8}}{1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_8 z^{-8}} \quad (9)$$

The transfer function (9) was approximated by the discrete second-order model with time-delay (1).

It is obvious that linear model (8) was obtained without complying with valves contain hysteresis and other nonlinearities that the series liquid tanks system contains (Chalupa et al. 2011).

IDENTIFICATION OF SERIES LIQUID TANKS

Determination of number time-delay steps

In this paper, the number of time-delay steps is obtained using an off-line identification by the least squares method (LSM). The measured process output (liquid level $h_8(k)$ [m] near operating flow) is influenced by input – generator of white noise which excites changes of flow rate $q_{1I}(k)$ [m³ min⁻¹]. The non-measurable system disturbances cause errors e in the determination of model parameters and therefore real output vector is in the form

$$y = F\Theta + e \quad (10)$$

The matrix F has dimension $(N-n-d, 2n)$, the vector y $(N-n-d)$ and the vector of parameter model estimates Θ $(2n)$. N is the number of samples of measured input and output data, n is the model order. It is possible to obtain the LSM expression for calculation of the vector of the parameter estimates

$$\hat{\Theta} = (F^T F)^{-1} F^T y \quad (11)$$

Equation (11), where $n = 8$, serves for calculation of the vector of the parameter estimates $\hat{\Theta}$ using N samples of measured input-output data. The form of

individual vectors and matrices in equations (10) and (11) are introduced in (Bobál et al. 2013a,b). Consider that model (1) is the deterministic part of the stochastic process described by the ARX (regression) model

$$y(k) = -a_1 y(k-1) - a_2 y(k-2) + b_1 y(k-1-d) + b_2 y(k-2-d) + e_s(k) \quad (12)$$

where $e_s(k)$ is the random non-measurable component.

The vector of parameter model estimates is computed by solving equation (11)

$$\hat{\theta}^T(k) = [\hat{a}_1 \quad \hat{a}_2 \quad \hat{b}_1 \quad \hat{b}_2] \quad (13)$$

and is used for computation of the predicted output

$$\hat{y}(k) = -\hat{a}_1 y(k-1) - \hat{a}_2 y(k-2) + \hat{b}_1 u(k-1-d) + \hat{b}_2 u(k-2-d) \quad (14)$$

The quality of identification can be considered according to error, i.e. the deviation

$$\hat{e}(k) = y(k) - \hat{y}(k) \quad (15)$$

Continuous-time system (8) was identified by discrete model (1) using off-line LSM (11) for different time-delay dT_0 ; $T_0 = 1$ min. The White Noise Generator was used as excitation input signal. A criterion of the identification quality is based on sum of squares of error

$$J_{\hat{e}_2}(d) = \sum_{k=1}^N \hat{e}^2(k) \quad (16)$$

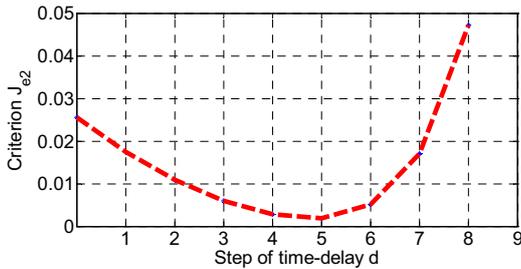


Figure 4: Criterion of Quality Identification for $d \in [0, 8]$

This criterion represents accuracy of process identification. It is obvious from Fig. 4 that minimum value of the criterion (21) is reached when the number of time-delay steps $d = 5$. Then it is possible to use model

$$\hat{G}_L(z^{-1}) = \frac{\hat{b}_1 z^{-1} + \hat{b}_2 z^{-2}}{1 + \hat{a}_1 z^{-1} + \hat{a}_2 z^{-2}} z^{-5} \quad (17)$$

for an approximation of model (8).

Identification procedures

Two identification procedures were used for calculation of parameter estimates of model (17). Following individual parameters were used for off-line LSM (11): $n = 2$; $d = 5$; $N = 300$.

Beside LSM the MATLAB function from the Optimization Toolbox

$$x = \text{fminsearch}('name_fce', x_0) \quad (18)$$

was also used for the off-line process identification. This function finds minimum of an unconstrained multivariable function using derivative-free method. Algorithm “fminsearch” uses the simplex search method of (Lagaris et al. 1998). This is a direct search method that does not use numerical or analytic gradients.

The difference between static gain $K_g = 3.08$ of the continuous-time transfer function (8) and estimation of the static gain of discrete transfer function (17) can serve as a good criterion for the quality of identification.

$$\hat{K}_g = \frac{\hat{b}_1 + \hat{b}_2}{1 + \hat{a}_1 + \hat{a}_2} \quad (19)$$

Identification using LSM

Discrete model for sampling period $T_0 = 1$ min

$$\hat{G}_{L1}(z^{-1}) = \frac{-0.0161z^{-1} + 0.0798z^{-2}}{1 - 1.7789z^{-1} + 0.7996z^{-2}} z^{-5} \quad (20)$$

was obtained using LSM method, $K_{g1} = 3.0733$.

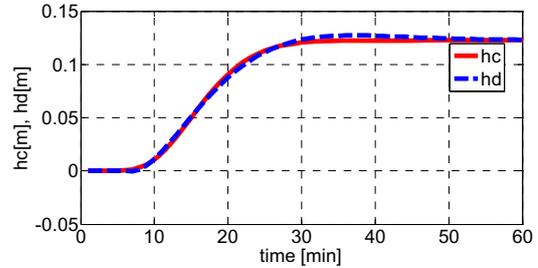


Figure 5: Comparison of step responses of models (8) and (20)

Comparison of step responses of continuous-time model (8) and discrete model (20) is shown in Fig. 5, where hc is the step response of the continuous-time model (8) and hd is step response of the discrete model (20). The input step signal $\Delta q_{1i} = 0.04 \text{ m}^3 \text{ min}^{-1}$ was chosen so that tank level is near to an operating point. It is obvious from numerator of the transfer function (20) than this system is slightly non-minimum phase (this is incurred by an identification error).

Identification using algorithm “fminsearch”

Discrete model for sampling period $T_0 = 1$ min

$$\hat{G}_{L2}(z^{-1}) = \frac{0.0309z^{-1} + 0.0286z^{-2}}{1 - 1.777z^{-1} + 0.7964z^{-2}} z^{-5} \quad (21)$$

was obtained using `fminsearch` method, $K_{g2} = 3.08$.

Comparison of step responses of continuous-time (8) and discrete model (21) is shown in Fig. 6. The input step signal Δq_{1l} is the same as in a previous case.

Model (21) is more accurate than model (20) and therefore was chosen for the design of two versions primary polynomial LQ controller for control of the series of liquid cylinder tanks.

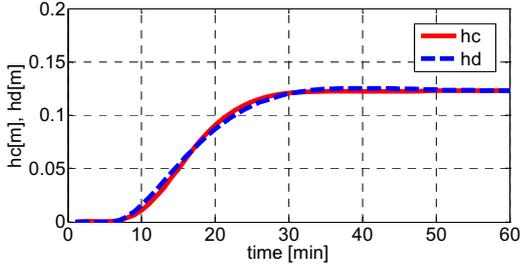


Figure 6: Comparison of step responses of models (8) and (21)

DESIGN OF PRIMARY POLYNOMIAL 2DOF LQ CONTROLLER

The design of the control algorithm is based on a general block scheme of a closed-loop with two degrees of freedom (2DOF) according to Fig. 7. The controller synthesis consists in the solving linear polynomial (Diophantine) equations. From first polynomial equation

$$A(z^{-1})K(z^{-1})P(z^{-1}) + B(z^{-1})Q(z^{-1}) = D(z^{-1}) \quad (22)$$

it is possible to compute 7 feedback controller parameters – coefficients of the polynomials Q , P . Polynomial $D(z^{-1})$ is the characteristic polynomial and $K(z^{-1}) = 1 - z^{-1}$.

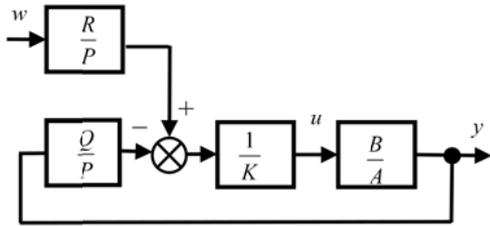


Figure 7: Block diagram of a closed loop 2DOF control system

Asymptotic tracking of the reference signal w is provided by the feedforward part of the controller which is given by solution of the following polynomial Diophantine equation

$$S(z^{-1})D_w(z^{-1}) + B(z^{-1})R(z^{-1}) = D(z^{-1}) \quad (23)$$

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 the controller design. Then it is possible to derive the polynomial R from equation (23) by substituting $z = 1$

$$R = r_0 = \frac{D(1)}{B(1)} \quad (24)$$

The 2DOF controller output is given by

$$u(k) = \frac{r_0}{K(z^{-1})P(z^{-1})} w(k) - \frac{Q(z^{-1})}{K(z^{-1})P(z^{-1})} y(k) \quad (25)$$

Two primary polynomial LQ controllers are derived in this paper using minimization of LQ criterion (Kučera 1991). Spectral factorization by means of the MATLAB Polynomial Toolbox 3.0 (Šebek 2014) is used for a minimization procedure.

The design of two LQ controllers for control of the second-order system with time-delay (1) is in detail derived in (Bobál et al. 2014; Bobál et al. 2015).

Minimization of LQ Criterion Using $u(k)$

In the first case the linear quadratic control methods try to minimize the quadratic criterion which uses penalization of the value of the controller output

$$J = \sum_{k=0}^{\infty} \left\{ [w(k) - y(k)]^2 + q_u [u(k)]^2 \right\} \quad (26)$$

where q_u is the so-called penalization constant, which gives the influence of the controller output to the value of the criterion. In this paper, criterion minimization (26) will be realized through the spectral factorization for an input-output description of the system

$$A(z)q_u A(z^{-1}) + B(z)B(z^{-1}) = D(z)\delta D(z^{-1}) \quad (27)$$

where δ is a constant chosen so that $d_0 = 1$. $A(z)$, $B(z)$ are the second-order polynomials and $D(z)$ is also the second-order polynomial

$$D(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} \quad (28)$$

Spectral factorization of polynomials of the first and the second degree can be computed by analytical way; the procedure for higher degrees must be performed iteratively (Bobál et al. 2005). The MATLAB Polynomial Toolbox is used for a computation of spectral factorization (27) using file `spf.m` by command

$$d = \text{spf}(a^*q_u^*a + b^*b) \quad (29)$$

It is known that by using the spectral factorization (27), it is possible to compute only two suitable poles (α , β). It is obvious from equation (22) that in this case a choice of the fourth-degree polynomial $D(z)$ is optimal

$$D_4(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + d_3 z^{-3} + d_4 z^{-4} \quad (30)$$

Therefore the other poles (γ , δ) are user-defined. A method for suitable pole assignment and computation of parameters of polynomial (30) was designed in

(Bobál et al. 2014). Then the primary digital 2DOF controller (25) can be expressed in the form

$$u(k) = r_0 w(k) - q_0 y(k) - q_1 y(k-1) - q_2 y(k-2) + (1-p_1)u(k-1) + p_1 u(k-2) \quad (31)$$

where

$$r_0 = \frac{1+d_1+d_2+d_3+d_4}{b_1+b_2} \quad (32)$$

and parameters q_0, q_1, q_2, p_1 are computed from (22).

Minimization of LQ Criterion Using $\Delta u(k)$

In the second case the linear quadratic control methods try to minimize the quadratic criterion which uses penalization of the incremental value of controller output

$$J = \sum_{k=0}^{\infty} \left\{ [w(k) - y(k)]^2 + q_u [\Delta u(k)]^2 \right\} \quad (33)$$

Equation (27) for computation of the spectral factorization changes into

$$(1-z)A(z)q_u(1-z^{-1})A(z^{-1}) + B(z)q_u B(z^{-1}) = D(z)\delta D(z^{-1}) \quad (34)$$

It is obvious that the characteristic polynomial in (34) is the three-degree polynomial

$$D(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + d_3 z^{-3} \quad (35)$$

Spectral factorization of (34) gives three optimal poles. However for the 2DOF controller design it is possible to propose other three user-defined real poles of the polynomial

$$D_6(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + d_3 z^{-3} + d_4 z^{-4} + d_5 z^{-5} + d_6 z^{-6} \quad (36)$$

The expressions for computation of individual parameters of polynomial (36) are derived in (Bobál et al. 2015).

Then the 2DOF controller design consists of determination of polynomial parameters (36) using command (29) from the Polynomial Toolbox and solution of the Diophantine equation for computation of feedback controller parameters

$$A_s(z^{-1})K(z^{-1})P(z^{-1}) + B(z^{-1})Q(z^{-1}) = D_6(z^{-1}) \quad (37)$$

where

$$A_s(z^{-1}) = 1 + a_{s1}z^{-1} + a_{s2}z^{-2} + a_{s3}z^{-3} \quad (38)$$

$$a_{s1} = a_1 - 1; \quad a_{s2} = a_2 - a_1; \quad a_{s3} = -a_3 \quad (39)$$

and

$$K(z^{-1}) = 1 - z^{-1}; \quad P(z^{-1}) = 1 + p_1 z^{-1} + p_2 z^{-2}; \quad (40)$$

$$Q(z^{-1}) = q_0 + q_1 z^{-1} + q_2 z^{-2} + q_3 z^{-3}$$

and from expression (24)

$$r_0 = \frac{1+d_1+d_2+d_3+d_4+d_5+d_6}{b_1+b_2} \quad (41)$$

The primary 2DOF controller output is given by

$$u(k) = r_0 w(k) - q_0 y(k) - q_1 y(k-1) - q_2 y(k-2) + (p_1 - p_2)u(k-2) - p_2 u(k-3) \quad (42)$$

SIMULATION VERIFICATION AND RESULTS

A simulation verification of the designed control algorithms was performed in MATLAB/SIMULINK environment. The robustness of individual control loops was experimentally investigated by a change of the static gain K of the nominal process model. From the point of view of the robust theory it is possible to consider these experiments as the gain margin determination by the parametric uncertainty influence. The experimental process model (8) was used for simulation experiments.

The individual simulation experiments are realized subsequently: the static gain $K_g = 3.08$ was increased as far as the control closed-loop was in the stability boundary. The experiments are not realized when the static $K_g = 3.08$ was decreased.

Control Using Primary Controller (31)

Because the subject of this paper is oriented to design of the polynomial robust control, the following simulation experiments have been realized. The discrete transfer function (21)

$$\hat{G}_{L2}(z^{-1}) = \frac{0.0309z^{-1} + 0.0286z^{-2}}{1 - 1.777z^{-1} + 0.7964z^{-2}} z^{-5} \quad (43)$$

with $K_{g2} = 3.08$ is the nominal model.

The penalization factor $q_u = 2$ was used for all experiments. The characteristic polynomial is given by

$$D_4(z) = z^4 - 3.3716z^3 + 4.4271z^2 - 2.4148z + 0.5154$$

with individual poles

$$\alpha, \beta = 0.8358 \pm 0.1301i; \quad \gamma = 0.8; \quad \delta = 0.9.$$

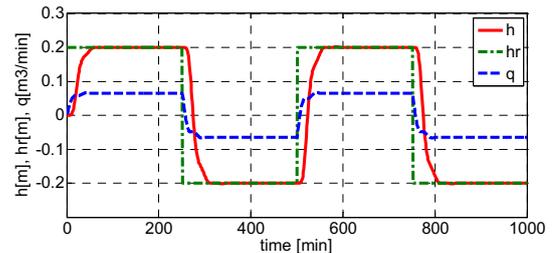


Figure 8: Control of nominal model $G_{L2}(z^{-1})$, $K_{g2} = 3.08$

The individual control parameters of controller (31):
 $q_0 = 0.8832$; $q_1 = -1.5652$; $q_2 = 0.6968$;
 $p_1 = -0.6218$; $r_0 = 0.0147$.

The control courses of the process output and controller output for the nominal model $G_{L2}(z^{-1})$ are shown in Fig. 8.

The discrete transfer function

$$\hat{G}_{P2}(z^{-1}) = \frac{0.0602z^{-1} + 0.0558z^{-2}}{1 - 1.777z^{-1} + 0.7964z^{-2}} z^{-5} \quad (44)$$

with $K_{g2c} = 1.95 * 3.08 = 6$ is the perturbed model when the closed-loop control is on the stability boundary. The control courses of the process output and controller output for perturbed model (44) are shown in Fig. 9.

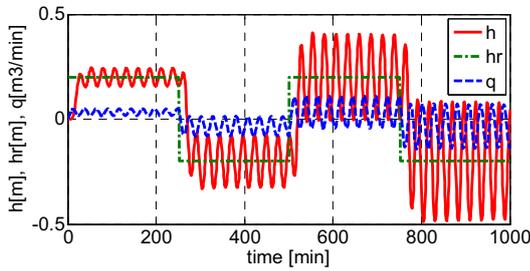


Figure 9: Control of perturbed model $G_{P2}(z^{-1})$,
 $K_{g2c} = 6$

It is obvious from Figs. 8 and 9 that approximate interval of the robust stability of nominal model $G_{L2}(z^{-1})$ by increase of the static gain is $K_{g2} \in (3.08, 6)$.

Control Using Primary Controller (42)

The discrete model (43) was chosen also as the nominal model. The characteristic polynomial is given by

$$D_6(z^{-1}) = 1 - 3.2798z^{-1} + 4.1025z^{-2} - 2.3413z^{-3} \\ + 0.5339z^{-4} - 0.0144z^{-5} + 0.0001z^{-6}$$

with individual poles

$$\alpha, \beta = 0.7925 \pm 0.25321i; \quad \gamma = 0.6948; \quad \delta = 0.01; \\ \lambda = 0.02; \quad \mu = 0.97.$$

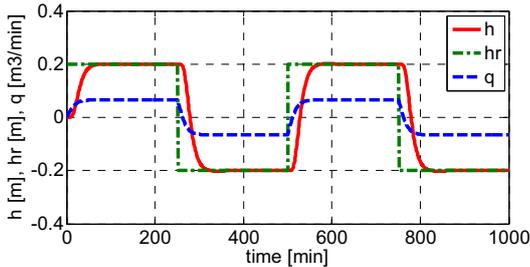


Figure 10: Control of nominal model $G_{DL2}(z^{-1})$,
 $K_{g2} = 3.08$

The individual control parameters of controller (42):

$$q_0 = 0.4891; \quad q_1 = -1.115; \quad q_2 = 0.8771; \quad q_3 = -0.2352; \\ p_1 = -0.0151; \quad p_2 = 1.9400e-04; \quad r_0 = 0.0160.$$

The control courses of the process output and controller output for the nominal model $G_{DL2}(z^{-1})$ are shown in Fig. 10.

The discrete transfer function

$$\hat{G}_{AP2}(z^{-1}) = \frac{0.0479z^{-1} + 0.0443z^{-2}}{1 - 1.777z^{-1} + 0.7964z^{-2}} z^{-5} \quad (45)$$

with $K_{g2c} = 1.55 * 3.08 = 4.77$ is the perturbed model when the closed-loop control is on the stability boundary. The control courses of the process output and controller output for perturbed model (45) are shown in Fig. 11.

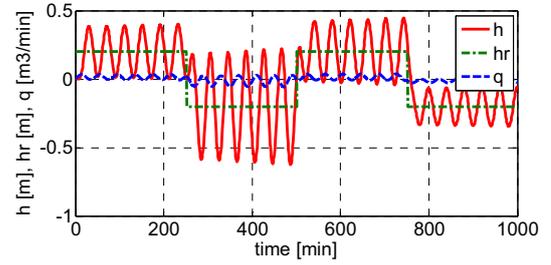


Figure 11: Control of perturbed model $G_{AP2}(z^{-1})$,
 $K_{g2c} = 4.77$

It is obvious from Figs. 10 and 11 that approximate interval of the robust stability of nominal model $G_{DL2}(z^{-1})$ by increase of the static gain is $K_{g2} \in (3.08, 4.77)$. It is possible to improve the robust stability of the controller (42) by increase of the penalization constant q_u . For example, when $q_u = 5$ the interval of the robust stability is $K_{g2} \in (3.08, 7.7)$.

CONCLUSION

Digital LQ Smith predictor algorithms for control of the high-order processes was designed. The high-order process, a set of equal liquid cylinder atmospheric tanks, was identified by the second-order model with five time- delay steps. The off-line least squares method was used for the identification of the number time-delay steps. The White Noise Generator was used as an excitation input signal. Two controller algorithms are based on polynomial design using the linear quadratic control method. This method minimizes the quadratic criterion by penalizing the value of the controller output $u(k)$ or its increment $\Delta u(k)$. The linear quadratic control method was combined with pole-assignment. Both designed controllers were derived to obtain algorithms with easy implementability in industrial practice. The control designs of both modifications were verified by simulation. The results of simulation verifications in both cases demonstrated very

good control quality and robustness of designed digital LQ algorithms. It is possible to improve the robust stability by increasing the penalization constant q_u . The contribution of this paper is the fact that a high-order system, which is composed of a set of low-order systems, can be approximated by a low-order model with time-delay. For these approximated model it is possible to design relatively simple digital controllers.

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.
- Bobál, V., Chalupa, P., Dostál, P. and M. Kubalčík. 2013a. "Identification and self-tuning predictive control of heat exchanger." In *Proceedings of the 2013 International Conference on Process Control*, Štrbské Pleso, Slovakia, 219-224.
- Bobál, V., Kubalčík, M., Dostál, P. and J. Matějčík. 2013b. "Adaptive predictive control of time-delay systems." *Computers and Mathematics with Applications*, 66, 165-176.
- Bobál, V., Chalupa, P., Dostál, P. and M. Kubalčík. 2014. "Digital control of unstable and integrating processes." *International Journal of Circuits, Systems and Signal Processing*, 8, 424-432.
- Bobál, V., Dostál, P., Kubalčík, M. and S. Talaš. 2015. „LQ control of heat exchanger – design and simulation study." In *Proc. of 29th European Conference on Modelling and Simulation*, Albena, Bulgaria, 239-245.
- Chalupa, P., Novák, J. and V. Bobál. 2011. "Modeling of hydraulic control valves". In *13th WSEAS International Conference on Automatic Control, Modelling and Simulation, ACMOS'11*, Lanzarote, Spain, 195-200.
- 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*. Instrument Society of America.
- Kučera, V. 1991. *Analysis and Design of Discrete Linear Control Systems*. Prentice-Hall, Englewood Cliffs, NJ.
- Kučera, V. 1993. "Diophantine equations in control – a survey". *Automatica* 29, 1361-1375.
- Lagaris, J. C., Reeds, J. A., Wright, M. H. and P. E. Wright. 1998. "Convergence properties of the Nelder-Mead simplex method in low dimensions." *SIAM Journal of Optimization*, 9, 112-147.
- Normey-Rico, J. E. and E. F. Camacho. 2007. *Control of Dead-time Processes*. Springer-Verlag, London.
- Smith, O.J. 1957. "Closed control of loops". *Chem. Eng. Progress* 53, 217-219.
- Šebek, M. *Polynomial Toolbox for MATLAB, Version 3.0*. 2014. PolyX, Prague, Czech Republic.
- Šebek, M. and M. Hromčík. 2007. „Polynomial design methods," *International Journal of of Robust and Nonlinear Control* 17, 679-681.

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. 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.



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.



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. His current work covers following areas: control of multivariable systems, self-tuning controllers, predictive control. His e-mail address is: kubalcik@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 Ph.D. study in the Department of Process Control. His e-mail address is talas@fai.utb.cz.

DISCRETE METHOD FOR ESTIMATION OF TIME-DELAY OUTSIDE OF SAMPLING PERIOD

Stanislav Talaš, Vladimír Bobál, Adam Krhovják and Lukáš Rušar
Tomas Bata University in Zlin
Department of Process Control, Faculty of Applied Informatics
T. G. Masaryka 5555
760 01 Zlin
Czech Republic
E-mail: talas@fai.utb.cz

KEYWORDS

Time-delay, on-line identification, time-varying delay, model based prediction, MATLAB/SIMULINK.

ABSTRACT

The aim of this paper is to suggest a new approach in time-delay identification. With the focus on the value of the delay it utilizes known system parameters for model based predictions in order to estimate the most precise value of the time-delay. This method was further extended by applying a modified internal model based on smaller sampling value which enables to receive results with greater precision than a single original sampling step. The suggested procedure therefore grants a deeper insight into the development of the time-delay and influencing effects from outer signals. The identification algorithm was realized in the MATLAB environment, which was also used for simulations verifying precision of the approach.

INTRODUCTION

Time-delay occurrence in industrial processes brings significant complications. As a cause of desynchronization between input and output signals it makes use of conventional control techniques ineffective. It is necessary to apply special control strategies designed specifically to deal with this kind of behaviour. Disadvantage of these control algorithms is, among others, an increase of demands for amount of recorded data and performed computations, as well as dependence on precise information about the duration of the delay (Normey-Rico and Camacho, 2007). Furthermore, number of control approaches which are sufficiently robust with respect to the time-delay is scarce (Normey-Rico and Camacho, 2008), (Karafyllis and Krstic, 2013). Therefore it is necessary for the value of the delay to be exactly determined and if possible even during the control of a process.

Due to the complexity of the time-delay identification problem and its resistance to conventional approaches, it remains a subject of active research (Richard, 2003). One of basic principles for time-delay estimation is based on calculation of correlation between input and output signals (Knapp and Carter, 2003). A number of

significant studies have also occurred in the area of relay based identification strategy working with oscillations and cycle measurements (Majhi, 2007). Another suitable strategy for adaptive algorithms is a multi-delay approach consisting of a number of possible delays with the goal of identifying the most precise one (Drakunov et al., 2006). Risk of changes in parameters during a control process suggests an application of recursive approach to identification. An extension of established algorithms with recursive delay estimation was already presented in (Elnaggar et al., 1989) In cases where time-delay may change in time, or depend on system parameters it is suitable to apply an adaptive identification technique. For such systems it can be assumed that the time-delay is the only unknown system parameter (Orlov et al., 2003). Adaptive or on-line identification strategies find practical applications by themselves as well as a part of a bigger adaptive system. A method of increased precision beyond limitations of sampling time based on least-squares method and frequency analysis was suggested in (Ferretti et al., 1991). Despite a number of scientific studies, the demand for precise time-delay estimation still remains.

With the intention to develop an on-line identification approach targeted exclusively on time-delay itself, we have proposed a multi-delay strategy derived from the predictive control. It is an identification technique focused purely on time-delay parameter with assumption that systems structure and remaining values are known

As the key mechanics the predictive equation and its defining parameters are analysed in the first chapter. Consequently, the proposed time-delay identification algorithm is described and illustrated step by step. The functionality verification follows in the next section in a form of simulated process with time varying time-delay parameter.

THEORETICAL BASIS FOR IDENTIFICATION

The goal during time-delay estimation is to find a particular shape of output signal corresponding to specific input signal. In order to enable the on-line application of the algorithm, it is necessary to address the fact that output shape can have almost any form. Even if we don't consider presence of noise, a number of ongoing output changes may exist, especially in systems of high-

er order. To safely assume the shape of the output in relation to the recorded input the knowledge of system parameters is necessary. Therefore, the design of the identification method assumes that the structure of the studied system and its parameters with the exception of time-delay itself are known.

Part of the design was the option to increase the precision through the application of system parameters to derive a mathematical description with lowered sampling period. The existence of multiple models led to suggestion of multi-delay approach with determining the most precise outcome based on qualitative criterion.

Predictive principle

The mechanism for estimating of future behaviour of the controlled system was chosen as the basic element for our identification technique. Its principle stands on using a mathematical model of the studied system as a simulated representative of the real system's behaviour. In its general form for 2nd order discrete system

$$G(z) = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \quad (1)$$

an equation for a future output estimation in declarative version can be stated (Haber et al., 2011)

$$\hat{y}(k+1) = -a_1 y(k) - a_2 y(k-1) + b_1 u(k) + b_2 u(k-1) \quad (2)$$

Provided that the control input development and the two past values of the output are known, it is possible to estimate the future outputs up to the duration of the control input.

$$\hat{y}(k+i) = -a_1 y(k-1+i) - a_2 y(k-2+i) + b_1 u(k-1+i) + b_2 u(k-2+i) \quad (3)$$

$$i = 1, 2, \dots, N$$

The amount of future steps which will be predicted is determined by constant N called prediction horizon. This value limits the length of predicted output, as well as computing demands of this operation.

Prediction with time-delay

In case of our method, we intend to estimate the output of a system containing time-delay, which the model may express in the following form

$$G(z) = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} z^{-d} \quad (4)$$

where d represents the number of sampling steps of time-delay. The previous relation (3) needs to be extended as

$$\hat{y}(k+i) = -a_1 y(k-1+i) - a_2 y(k-2+i) + b_1 u(k-1+i-d) + b_2 u(k-2+i-d) \quad (5)$$

$$i = 1, 2, \dots, N$$

Data from the record of the input signal are then shifted by the length determined by the delay duration.

METHOD OF ESTIMATION

Based on knowledge of remaining system parameters, it is possible to determine the output values in the future sampling steps. When we apply this procedure for a series of internal models with a value of time-delay varying in a specified interval a number of possible outcomes is obtained. Consequently, these results are compared with measured output and the most precise set gives the most probable value of the time-delay.

In order to eliminate the necessity to wait the length of the prediction horizon to receive all the data for comparison the whole procedure was transformed to work only with data already measured. Initial values for the prediction are therefore extracted from the output data measured at horizons length in the past, as well as control input values.

$$\begin{aligned} \hat{y}(k-i) &= -a_1 y(k-1-i) - a_2 y(k-2-i) \\ &\quad + b_1 u(k-1-i-d) + b_2 u(k-2-i-d) \quad (6) \\ i &= N-1, N-2, \dots, 0 \end{aligned}$$

For illustration purposes the following data are from simulated identification of a 2nd order linear system with parameters

$$\frac{2}{4s^2 + 5s + 1} \quad (7)$$

in its discrete form

$$\frac{0,4728z^{-1} + 0,2076z^{-2}}{1 - 0,7419z^{-1} + 0,0821z^{-2}} \quad (8)$$

with sampling period $T_0 = 2s$ and time-delay d set to 4.5s. The prediction horizon size was set to 10 sampling steps.

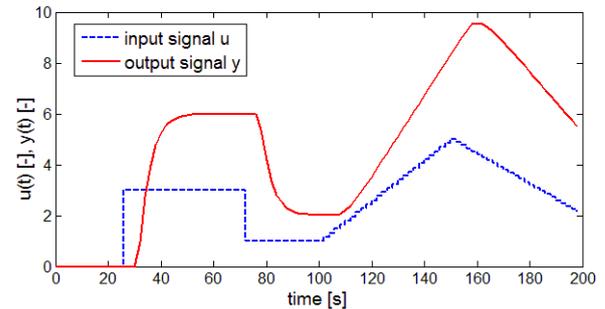


Figure 1: Controlled process as a source of input and output values for identification

Applying the identification procedure at time of 90s with maximal expected delay set to 12s provides the following range of possible outcomes.

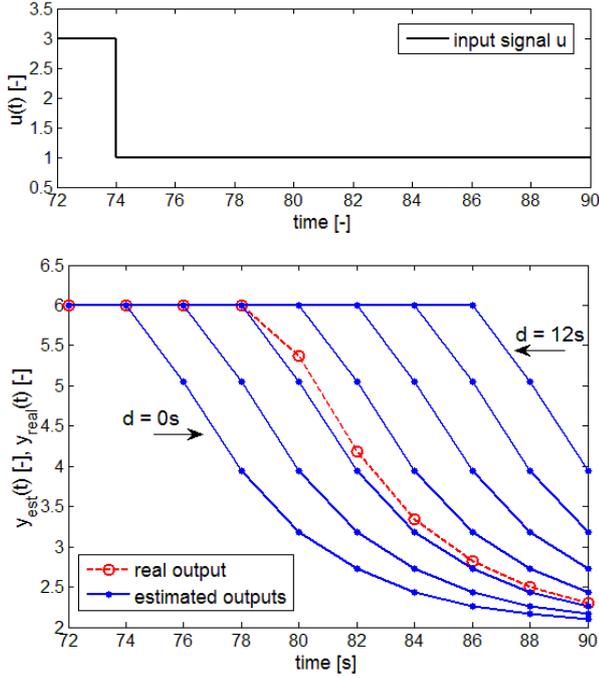


Figure 2: Set of possible system outputs based on different time-delay values

Estimated outputs are compared with the real one and using a discrete version of a qualitative criterion ISE its dependency on the duration of the time-delay is obtained.

$$ISE(k, d) = \sum_{i=0}^{N-1} [y_d(k-i) - y(k-i)]^2 \quad (9)$$

where k represents the current step. Therefore we obtain a set of values representing divergences from delay estimations in every sampling step.

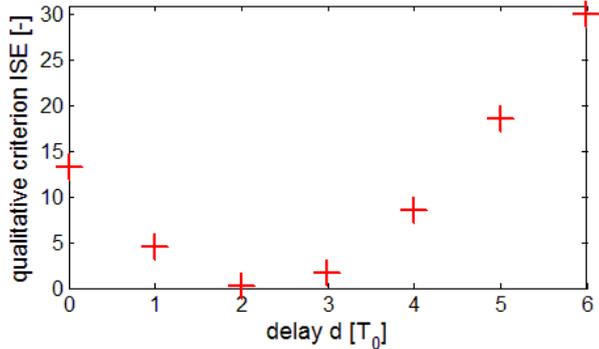


Figure 3: ISE criterion dependency on the estimated value of time-delay

Considering the content of Figure 3 we may assume, that data interpolation followed by determining its minimal value may produce even more precise estimation. Nonetheless, interpolation cannot guarantee notably precise information about the development of the system dynamics in reference to the time-delay. Therefore, the internal model is modified to a form with decreased sampling period. This model is again applied in the identification algorithm. The control input signal has to be shaped accordingly to fit the new sampling time. In

order to determine initial values of the output an interpolation between two recorded values was used. To decrease unnecessary computations, this time the predictions are performed only in the interval between two smallest values gained by the criterion in Figure 3.

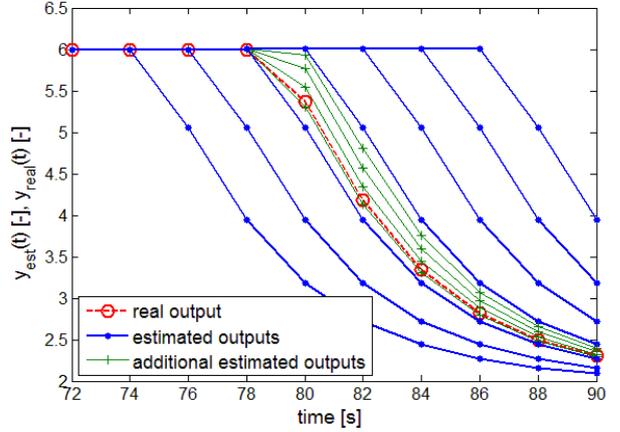


Figure 4: Extended set of possible system outputs

Repeated evaluation of the precision by applying the identical criterion (9) for the newly received estimates offers the result with a greater amount of data in the critical area.

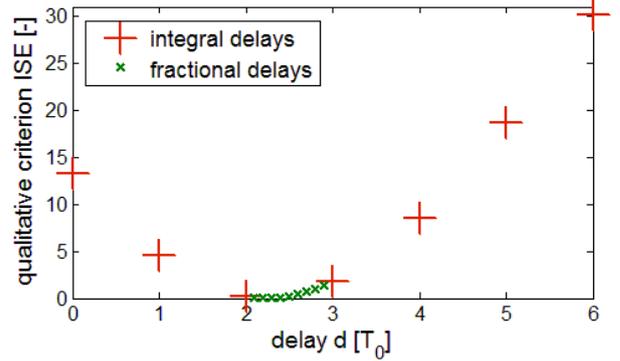


Figure 5: Extended ISE criterion dependency on the estimated value of time-delay

Figure 5 demonstrates how additional data received from the modified internal model with sampling time 0.2s fill in the missing part of the crucial segment.

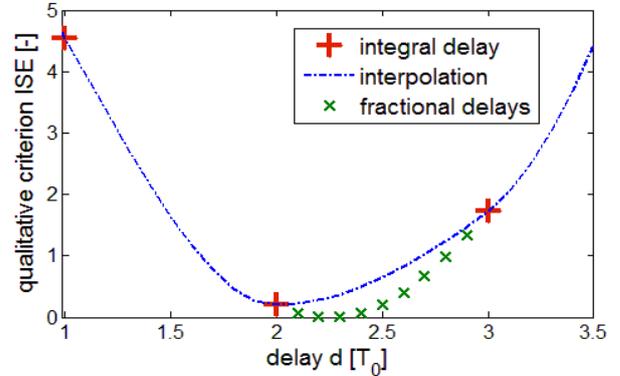


Figure 6: Comparison between identification results and interpolation

The distinction from the interpolated data is displayed in Figure 6, where piecewise cubic Hermite interpolation

and our estimation results were compared. A significant difference appears between both minimal values. Length of the horizon determines the amount of data used in prediction and therefore criterion calculations. Larger area of identification can more accurately determine the correct outcome especially if noise is present. Nevertheless, during a sudden change in time-delay, in theory, approximately half of the horizon length of the new data needs to be processed in order to establish a new minimum in the criterion set. Therefore, changing the horizon value shifts the balance between precision and speed of the identification.

ALGORITHM VERIFICATION RESULTS

The identification algorithm performance was tested by a simulated system with time-varying delay. System parameters were kept identical to (7) (8), only the value of the delay was set to change. The length of the prediction horizon was set to 10 sampling steps which were previously verified to provide acceptable results. The procedure was applied in every sampling step, however several first steps could not be analysed as the amount of data was not sufficient to fill the prediction horizon. Ideal conditions for identification require constant changes in recorded signals. Therefore, the shape of control input consists of ramps and step changes.

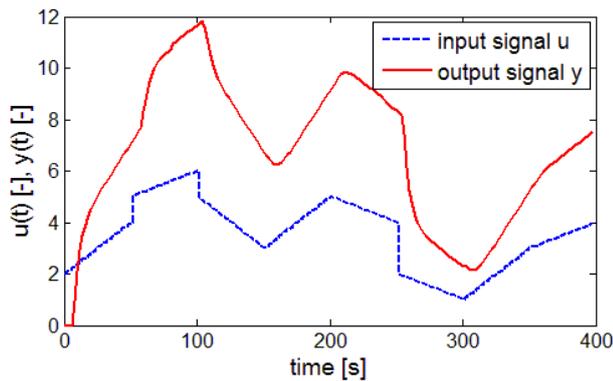


Figure 7: Controlled process as a source of input and output values for identification

Figure 7 shows recorded input and output signals which were used in the identification algorithm.

To test the option of on-line application a time-varying output signal delay was suggested. Its final form contains areas of constant value, sudden steps and ramps.

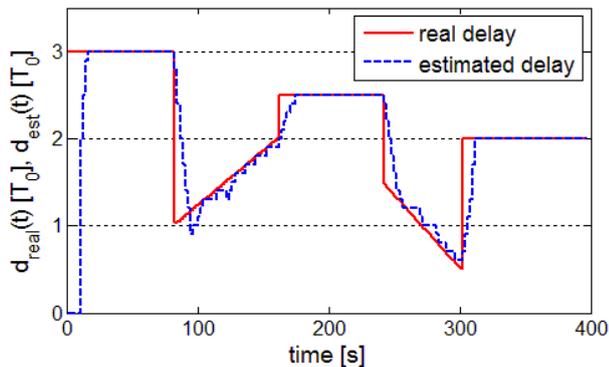


Figure 8: Time-delay on-line estimation

As can be seen in the Figure 8 delay estimation results are able to determine the correct value of the system time-delay while it is static. To increase the precision during large step changes of the time-delay, the maximal change in results has been limited to a half of the sampling period. Consequently, the development of the estimations is slower, but on the other hand outcomes are closer to real values. Constant changes in the form of ramps have proven to be the most difficult to estimate, for the mere fact that the bigger is the delay the longer it takes to get the correct result, which tends to distort results in the whole section of the process.

Noise compensation

From the principle of the method is clear, that the outcome precision depends on how the real process behaviour is close to the internal model description. We have selected noise as a tool to simulate inaccuracies of a real system.

To demonstrate the negative influence of noise, the previous model (7) was extended with a source of white noise. The input signal trajectory was kept identical, as well as the time-delay development.

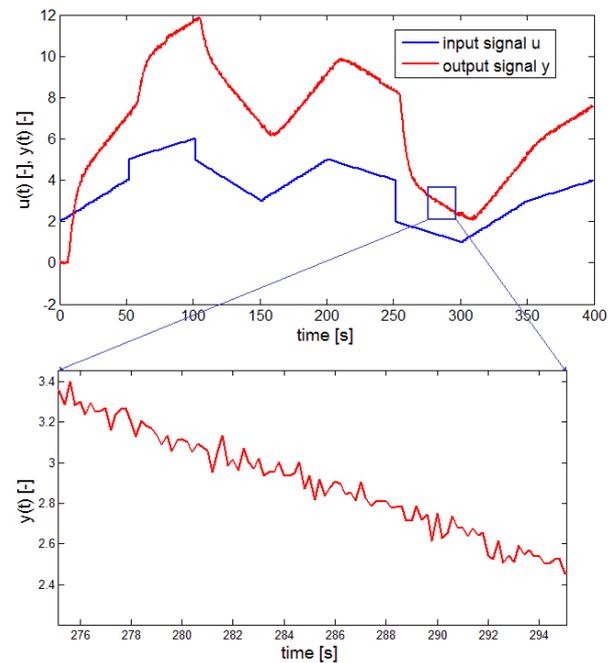


Figure 9: Controlled process with added noise

Figure 9 displays the system input signal and delayed output burdened by noise. This process was initially identified with algorithm parameters identical to the previous case.

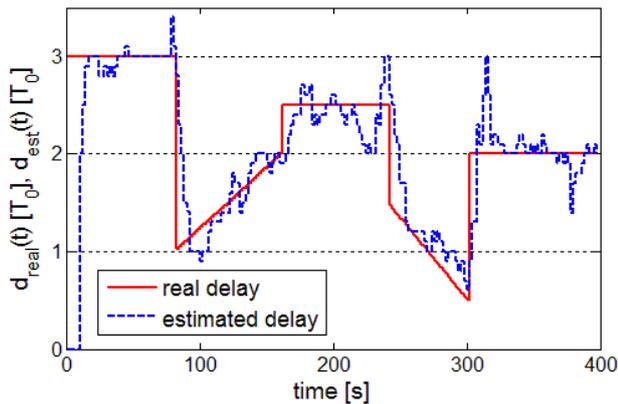


Figure 10: Time-delay online estimation from process with added noise

Figure 10 shows that noise can significantly influence the value of qualitative criterion and consequently entire identification result. Noise influence heavily depends on the ratio between noise and system gain, as well as chosen precision of the method.

In order to compensate above mentioned negative effects, we have expanded the prediction horizon, which has enlarged the area involved in determining of the estimation precision. Therefore, the prediction horizon N in the identification algorithm was set to 15 sampling steps in order to decrease the influence of the noise, which has extended the area studied in every step by 10 seconds of recorded data.

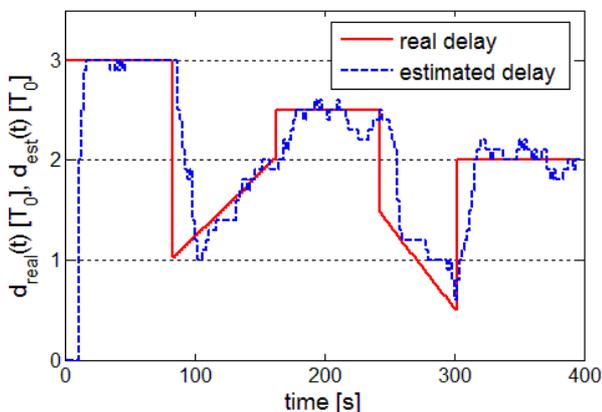


Figure 11: Time-delay on-line estimation from process with added noise and greater value of the prediction horizon

Figure 11 demonstrates the effectiveness of widening of the prediction horizon as a significant amount of considerably differing results has been cleared away. On the other hand, due to the higher value of the prediction horizon, sections with constant change of time-delay tend to be processed less accurately and expressed in small periodic steps. Nevertheless, the overall difference between true and estimated delay has been diminished, therefore it seems safe to claim that the bigger prediction horizon is able to a certain extent negate effects of imprecision.

CONCLUSION

In this paper we have suggested a new discrete method for time-delay estimation, enabling to determine its value even in cases when it is not an integer. Extending of the multi-delay approach with a version of the studied system with a smaller sampling step made possible to predict output values of system with a fractional time-delay. Functionality of the designed method was verified in simulations and has proven the correctness of its results.

Periodic measurements during a process control can provide an opportunity to measure changes in time-delay effect during the process that would stay otherwise hidden. Additionally, they provide information about a possible influence of system states on time-delay. The estimation concept can be applied both on-line and outside of a controlled process. It is also suitable for any type of system which can be described in a form of internal model.

Further research involves application of the method on a real system and increase in robustness of the identification algorithm.

ACKNOWLEDGEMENT

This article was supported by Internal Grant Agency of Tomas Bata University under the project No. IGA/FAI/2016/006.

REFERENCES

- Drakunov, S. V., W. Perruquetti, J.-P. Richard and L. Belkoura (2006). Delay identification in time-delay systems using variable structure observers. *Annual Reviews in Control*, **30**, 143-158.
- Elnaggar, A., G. A. Dumont and A.-L. Elshafei. (1989). Recursive estimation for system of unknown delay. *Proceedings of 28th Conference on Decision and Control*, 1809-1810.
- Ferretti, G., C. Maffezzoni and R. Scattolini (1991). Recursive estimation of time delay in sampled systems. *Automatica*, **27** (6), 653-661.
- Haber, R., R. Bars and U. Schmitz, (2011). Predictive control in process engineering: From basics to the applications. Willey-VCH Verlag, Weinham.
- Karafyllis I. and M. Krstic (2013). Delay-robustness of linear predictor feedback without restriction on delay rate, *Automatica*, **49** (6), 1761-1767.
- Knapp, C. and C. C. Carter (2003). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing*, **24** (4), 320-327.
- Majhi, S. (2007). Relay based identification of processes with time delay. *Journal of Process Control*, **17** (2), 93-101.
- Normey-Rico J. E. and E. F. Camacho (2007). Control of dead-time processes. London, Springer-Verlag.
- Normey-Rico J. E. and E. F. Camacho (2008). Dead-time compensators: A survey. *Control in Engineering Practice*, **16**, 407-428.

Orlov, Y., L. Belkoura, J. P. Richard and M. Dambrine (2003). Adaptive identification of linear time-delay systems. *International Journal of Robust and Non-linear Control*, **13** (9), 857-872.

Richard, J.P. (2003). Time-delay systems: an overview of some recent advances and open problems. *Automatica*, **39** (10), 1667-1694.

AUHOR 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 interests focus on modelling and simulation of continuous time technological processes, adaptive and nonlinear control. He is currently working on programming simulation library of technological systems.

LUKÁŠ RUŠAR studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2014. 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 focuses on model predictive control. His e-mail address is rusar@fai.utb.cz.

NONLINEAR SIMULINK MODEL OF MAGNETIC LEVITATION LABORATORY PLANT

Petr Chalupa
Martin Malý
Jakub Novák

Faculty of Applied Informatics
Tomas Bata University in Zlin
nam. T. G. Masaryka 5555, 760 01, Czech Republic
E-mail: chalupa@fai.utb.cz

KEYWORDS

First principle modelling, MATLAB, Simulink, magnetic levitation, CE152 model

ABSTRACT

The paper deals with modelling of a magnetic levitation laboratory plant. The goal of the work was to create a nonlinear model in a MATLAB / Simulink environment representing behaviour of a real-time CE152 laboratory plant. The CE152 is a magnetic levitation model developed by Humusoft company. From the control point of view, the CE152 magnetic levitation plant is a nonlinear very fast system. The model of the plant is developed using first principle modelling and subsequently made more precise using real-time experiments. The behaviour of the resulting Simulink model is compared with the behaviour real-time plant. The Simulink model can be further used in the process of controller design.

INTRODUCTION

Knowledge of a model a controlled plant is necessary for most of current control algorithms (Bobál et al. 2005). It is obvious that some information about controlled plant is required to allow design of a controller with satisfactory performance. A plant model can be also used to investigate properties and behaviour of the modelled plant without a risk of damage of violating technological constraints of the real plant. Two basic approaches of obtaining plant model exist: the black box approach and the first principles modelling (mathematical-physical analysis of the plant).

The black box approach to the modelling (Liu 2001), (Ljung 1999) is based on analysis of input and output signals of the plant. In this case the knowledge of physical principle of controlled plant is not required but obtained model is generally valid only for signals it was calculated from.

The first principle modelling provides general model valid for whole range of plant inputs and states. The model is created by analysing the modelled plant and combining physical laws (Himmelblau and Riggs 2004). On the other hand, there are usually many unknown constants and relations when performing analysis of a plant. Therefore, modelling by first principle modelling

is suitable for simple controlled plants with small number of parameters. First principle modelling can be used for obtaining basic information about controlled plant (range of gain, rank of suitable sample time, etc.). Some simplifications must be used to obtain reasonable results in more complicated cases. These simplifications must relate with the purpose of the model. The first principle modelling is also referred to as white box modelling.

The paper combines of both methods. Basic relations are derived using first principles. The obtained model is further improved on the basis of measurements. This approach is known as grey box modelling (Tan and Li 2002). The goal of the work was to obtain a mathematical model of the CE152 Magnetic levitation plant (Humusoft 1996) and to design the model in MATLAB-Simulink environment. The CE152 plant was developed by Humusof Ltd. and serves as a real-time model of fast nonlinear system. The major reason for creating the model of this laboratory equipment was usage of the model in control design process.

THE CE152 MAGNETIC LEVITATION SYSTEM SPECIFICATION

A photo of the CE 152 magnetic levitation plant is presented in Figure 1.



Figure 1: Photo of the CE152 plant

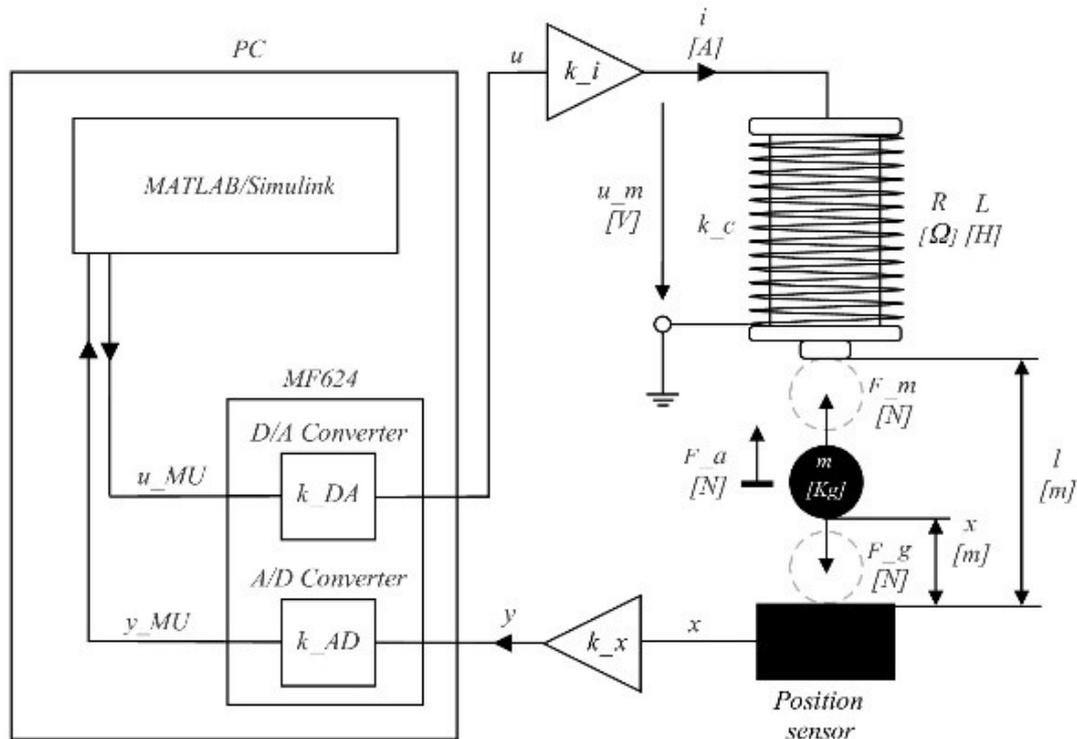


Figure 2: Principal scheme of the magnetic levitation model

The CE152 Magnetic levitation system is a nonlinear unstable dynamic system with one input and one output. Input signal is the control signal and output signal gives information about position of a steel ball. Both signal values are converted and scaled to the specific range of the machine unit [MU].

Structure of the CE152 Magnetic levitation system

The system consists of a model of the magnetic levitation system, power supply and a universal data acquisition card MF624. MF624 is a standard PCI card with A/D, D/A converters, analogue/digital inputs and outputs, counters, timers and appropriate drivers. The model is connected to the PC via this card.

Following parts are considered for a modelling of the plant:

- D/A converter,
- A/D converter,
- the position sensor,
- the power amplifier,
- the ball and coil subsystem

Simplified inner structure is shown in Figure 2. A steel ball levitates in magnetic field of the coil driven by power amplifier connected to D/A converter. Position of the steel ball is measured by inductive linear position sensor connected to A/D converter. Both control and measured parameters are sent and received by Simulink.

System behaviour

In this part system behaviour is discussed. As mentioned before the CE152 Magnetic levitation system is a nonlinear unstable dynamic system with one input and one output. When an input control signal of certain value is sent to the system, the ball is lifted upwards to the magnetic core and it stops when it hits the core. This behaviour is caused by electromagnetic force of the magnetic core which overcame the force of gravity. As the ball getting closer to the coil core, accelerating force grows. Because of an obstacle in the form of magnetic core, both the ball and accelerating force stops. Higher input signal means higher electromagnetic force and much more rapidly increasing acceleration. If input control value is decreased under the certain value, the ball falls down. That is happening, because the electromagnetic force is too low to overcome gravitational force. Ball fell down to the head of the sensor, which stops the ball. When the ball hits the sensor it bounces a several times. Presented behaviour is the system basic behaviour as a reaction to a constant input signal. Figure 3 shows the step response on different input values and Figure 4 presents the reaction, when the input value has changed to zero. If input value is not exactly zero, but still too low to hold up the ball, then the ball still falls, but his bouncing behaviour is slightly different due to the non-zero force of attraction.

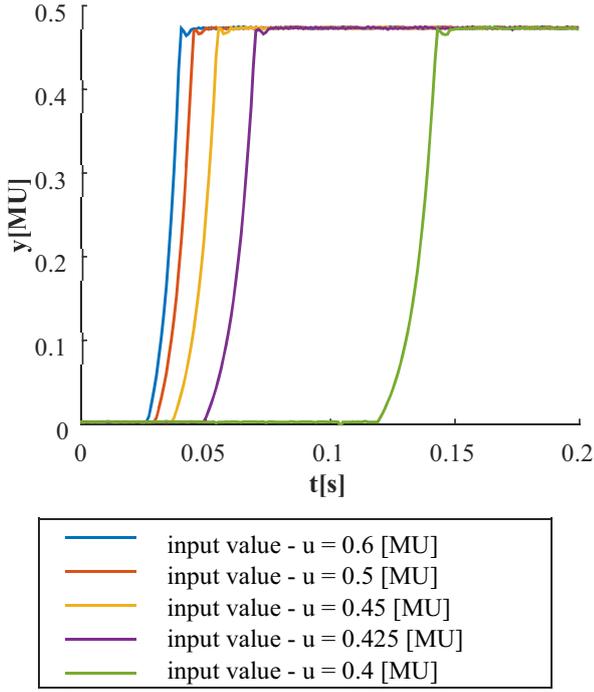


Figure 3: Step response on different input value

Tested input signal was step signal because other input signals like ramp or sinus signal don't have adequate information value, thanks to the described system behaviour, when once the electromagnetic force is strong enough, the ball is attracted to magnetic core, despite increasing input signal. In opposite case, when input signal decreasing, it is the same situation but in opposite meaning of the ball movement.

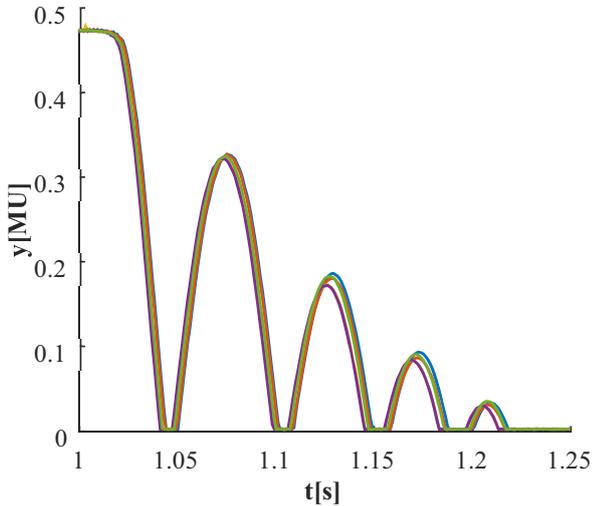


Figure 4: Reaction on input value change to zero

As it can be seen, there is some kind of delay, before the system starts to attract the ball. This is happening because coil first needs to attract the ball straight to the middle under its core and then the ball is lifted. This can be overcome by setting an optimal starting point of the ball. Even though, the small but non zero delay still remains and it is probably caused by internal processes in the system. Because of later comparison of the created model

with the real system, counting with this delay should be taken into account. A small delay can be also observed when the ball rebound (see Figure 4). This behavior is caused by difference between zero position of the measurement system and position corresponding to the ball on the head of the position sensor (ball in the initial position).

MODELLING OF THE SYSTEM PARTS

Based on mentioned approaches and methodologies, an adequate model of CE152 magnetic levitation system is deduced in this section. According to above preview of the system specification and composition, model can be decomposed to following parts: D/A and A/D converter, position sensor, power amplifier and ball and coil subsystem. All parts can be identified separately. All parts are eventually interconnected and they form the final model of CE152 magnetic levitation system.

D/A converter

The D/A converter converts digital signal u_{MU} from PC into an analogue voltage signal u . The D/A converter can be described by a linear function (1) and represented by a Simulink block (Figure 5).

$$u = k_{DA}u_{MU} \quad (1)$$

where u is D/A converter output signal/coil input voltage [V], u_{MU} is D/A converter input signal [MU] and k_{DA} is D/A converter gain [V/MU]

The D/A converter maps the input signal range $u_{MU} < -1 \text{ MU}, 1 \text{ MU} >$ to the range $u < 0 \text{ V}, 5 \text{ V} >$. This leads to converter gain $k_{DA} = 10 \text{ V/MU}$. Real range of D/A input signal is -10 V to 10 V but input to the CE152 magnetic levitation system is constructed for the range 0 V to 5 V , so signal input must be constrained.

A/D converter

The A/D converter converts analogue voltage signal y into a digital signal y_{MU} . The A/D converter can be described by a linear function (2) and represented by a Simulink block (Figure 6).

$$y_{MU} = k_{AD}y \quad (2)$$

where y represents A/D converter output signal/position sensor voltage [V], y_{MU} is A/D converter output signal [MU] and k_{AD} is A/D converter gain [MU/V]

The A/D converter maps the input signal range $y < -10 \text{ V}, 10 \text{ V} >$ to the range $y_{MU} < -1 \text{ MU}, 1 \text{ MU} >$. This leads to converter gain $k_{AD} = 0.1 \text{ MU/V}$.

The position sensor

An inductive position sensor is used to measure the ball position x . Maximum declared height l is calculated as a difference between physical distance of the coil and the sensor l_0 and the ball diameter d_k . The position of the ball is obtained by reading the voltage from A/D converter's

output. Sensor voltage varies with ball position. The relation between ball position and voltage is approximately linear.

$$y = k_x x + y_0 \quad (3)$$

where x is ball position [m], y_0 position sensor offset [V], y position sensor voltage [V], k_x - position sensor gain [V/m]

Calibration experiment must be done. First of all, a travelling distance of the ball l has to be calculated. It is a difference between physical distance of the coil and the sensor $l_0 = 18.4 \cdot 10^{-3} m$ and the ball diameter $d_k = 12.7 \cdot 10^{-3} m$.

$$l = l_0 - d_k = 5.7 \cdot 10^{-3} m$$

Final results of the boundary values were noted (Table 1), and based on calculation with these values, position sensor gain is obtained.

Table 1: Calibration data of the position sensor

i	$x_i[m]$	$y_{MUi}[mV]$	$y_i[V]$
1	0	0.00254	0.02537
2	0.0057	0.47384	4.73840

Position sensor offset y_0 is taken as an initial value from the inductive sensor, when input action signal is zero.

$$y_0 = y_1 = 0.02537 V \quad (4)$$

$$k_x = \frac{y_2 - y_1}{x_2 - x_1} = 826.8525 V/m \quad (5)$$

The power amplifier

The power amplifier works as transconductance amplifier whose differential voltage between input voltage u from the D/A converter and u_i produces an output current supplied to the coil. The power amplifier essentially represents a source of constant current with the current stabilisation. Internal structure of the amplifier is presented in Figure 5 where k_{am} stands for amplifier gain [-] k_s stands for current sensor gain [-], L represents coil inductance [H], R_c is coil resistance [Ω] and R_s is resistance of feedback resistor [Ω]

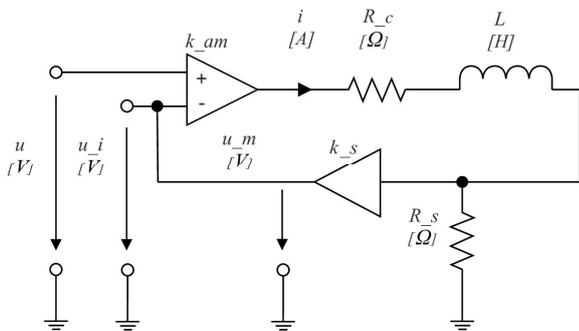


Figure 5: Internal structure of the power amplifier

The power amplifier can be described by the following set of equations:

$$u_m = \frac{di}{dt} L + i(R_c + R_s) \quad (6)$$

$$u_m = k_{am}(u - R_s k_s i) \quad (7)$$

Using direct Laplace transform with zero initial conditions leads to:

$$\frac{I(s)}{U(s)} = \frac{\frac{k_{am}}{R_c + R_s + k_{am} R_s k_s}}{\frac{L}{R_c + R_s + k_{am} R_s k_s} s + 1} \quad (8)$$

Simulation diagram is then shown in Figure 9.

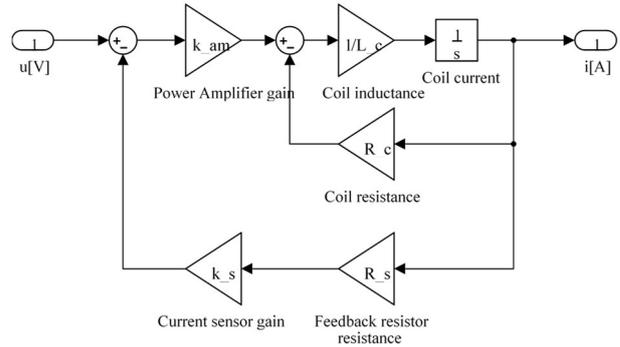


Figure 6: Power amplifier detailed simulation diagram

We can also do simplification of equation (8) and define it by transfer function of 1st order:

$$G_{PA}(s) = \frac{I(s)}{U(s)} = \frac{k_i}{T_a s + 1} \quad (9)$$

Constant k_i is an amplifier gain and T_a is an amplifier time constant. Manual (Humusoft 1996) states, that typical parameters of each component of the coil and power amplifier are:

$$k_{am} = 100$$

$$k_s = 13.33$$

$$L = 0.03 H$$

$$R_c = 3.5 \Omega$$

$$R_s = 0.25 \Omega$$

When these values are substituted into equations (8) and (9) then power amplifier gain and time constant have following values:

$$T_a = 8.90210^{-5} s$$

$$k_i = 0.297 A/V$$

Time constant is very small and can be neglected, then only power amplifier gain can remain. Difference between response of the detailed model and the simplified model with only power amplifier gain was assessed as minimal. Because of this minimal difference and faster computation of a simulation, simplified model was used.

The ball and coil subsystem

Lagrange's method can be used for modelling ball and coil subsystem. The motion equation is based on the balance of all acting forces. Final equation (11) is a nonlinear second order differential equation where input variable is electric current and output variable is ball position. Gravitational force F_g depends on ball mass and it is constant. This force acts against electromagnetic force F_m created by coil, when electric current pass through it. It is clear that to lift the ball up, electromagnetic force must be greater than gravitational force (accelerating force greater than zero). From the equation (11) we can figure out that electromagnetic force relies on two parameters, the amount of electric current i and actual ball position x . Damping force is also considered in this model:

$$F_a = F_m - F_g \quad (10)$$

$$m_k \ddot{x} + k_{fv} \dot{x} = \frac{i^2 k_c}{(x-x_0)^2} - m_k g \quad (11)$$

where:

F_m - electromagnetic force [N]

F_a - accelerating force [N]

F_g - gravitational force [N]

g - gravitational acceleration [$m \cdot s^{-2}$]

x - ball position [m]

m_k - ball mass [kg]

x_0 - coil offset [m]

k_c - coil constant [-]

i - coil current [A]

k_{fv} - dumping constant [$N \cdot s \cdot m^{-1}$]

Parameter of the ball mass m_k can be calculated from its diameter $d_k = 12.7 \cdot 10^{-3} m$ and density $\rho = 7800 kg \cdot m^{-3}$:

$$m_k = \rho V_k = \rho \frac{4}{3} \pi \left(\frac{d_k}{2}\right)^3 = 8.37 \cdot 10^{-3} kg \quad (12)$$

Parameters of the coil constant are obtained through the measurement of the steel ball stable positions, which means that the system have to be controlled somehow. Required data was obtained from closed loop control. Design of the control was taken over from Humusoft real time toolbox control simulation model for CE152 magnetic levitation system. Calculation of desired parameters has been done by two point calibration method with data given by (Table 2).

Table 2 Two point calibration data

i	y_{MUi}^s [MU]	x_i^s [m]	i_i^s [A]	u_i^s [V]
1	0.1500	0.018	0.6579	2.2152
2	0.3500	0.0042	0.4144	1.3951

Two stable points were measured and based on the following equation desired parameters were calculated.

$$x_0 = 0.0083[m];$$

$$k_c = 8.0915 \cdot 10^{-6}[N \cdot m^2 A^{-2}]$$

Parameter of the ball damping constant k_{fv} cannot be measured directly or by a dedicated experiment. It was identified by the trial-and-error method using real experimental data for comparison:

$$k_{fv} = 0.0195[N \cdot s \cdot m^{-1}]$$

The ball and coil simulation diagram is presented in Figure 7.

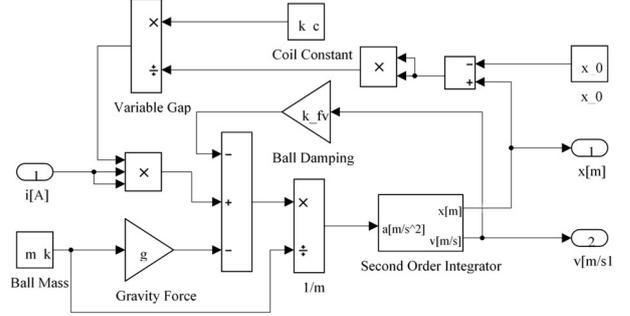


Figure 7: The ball and coil simulation diagram

Bouncing

A coefficient of restitution is used for bouncing ball behaviour. This coefficient accounts for the loss of kinetic energy during each bounce. Bouncing behaviour of the model isn't very close to the behaviour of the real system, because it doesn't take into account friction and other impacts of physics. Comparing can be found in section dedicated to validation and verification of the model. Bouncing model in placed inside the Second order integrator block presented in Figure 7. Bouncing subsystem itself is presented in Figure 8.

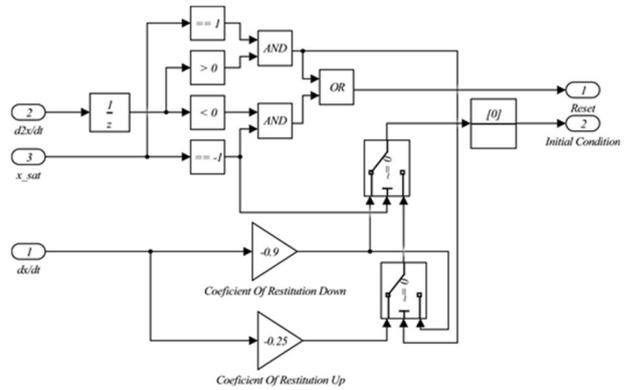


Figure 8: Hard stop and bouncing block diagram

THE WHOLE SYSTEM MODEL

The whole system consists of created subsystem, which were included into subsystem blocks and interconnected. Interconnection of these blocks is shown in Figure 9.

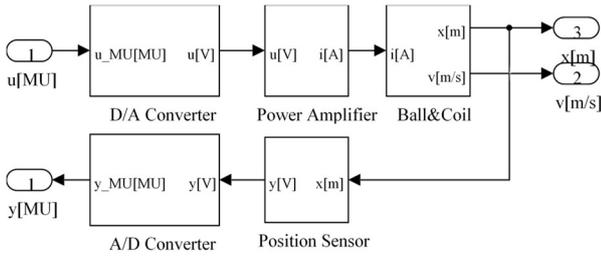
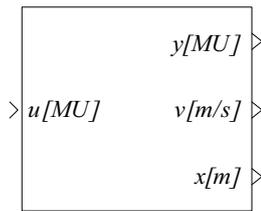


Figure 9: Interconnection of subsystem blocks

Finally, subsystem blocks were turned into a parent subsystem block as presented in Figure 10.



Nonlinear Magnetic Levitation Model

Figure 10: CE152 magnetic levitation system

VERIFICATION AND VALIDATION OF THE MODEL

From the perspective of a verification, created model is satisfactory. Behaviour of the model is just like the behaviour of the real system template. When we take a look on results of a validation, they are surprisingly good. The control input values have the same course and same is it with output values. The resulting characteristics depends strictly on estimated parameters. Figure 11 presents the response of the created model of the system to the input value $u = 0.45 \text{ MU}$ and its confrontation with response of the real system.

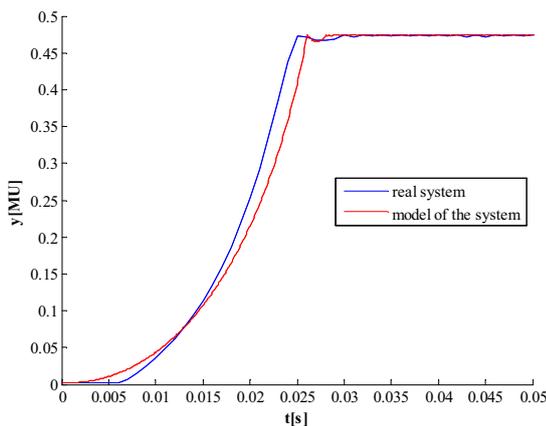


Figure 11: Comparison of step responses of the model and the real system

Response of the real system is then moved a little backwards, because of the delay discussed in a chapter dedicated to the system behaviour. From the same reason also response on the change of input value to zero needs to be moved a little backwards. This corresponds to an

immediate response of the system. Now if responses are compared it can be said, that created model have the same behaviour. Waveforms have the same direction and tendency.

Bouncing behaviour of the model is presented in Figure 12 and differs from the behaviour of the real plant. But this is obvious, because this behaviour was modelled in the simplest form.

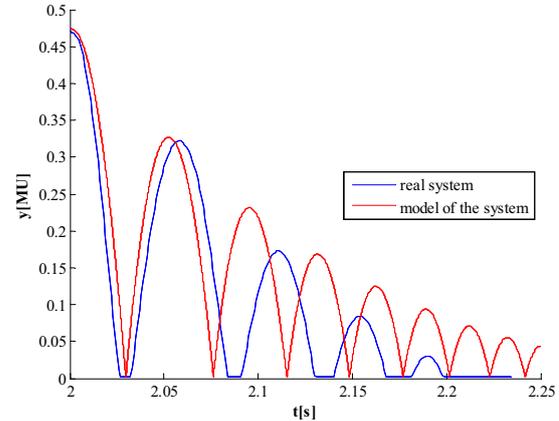


Figure 12: Comparison of bouncing of the model and the real system

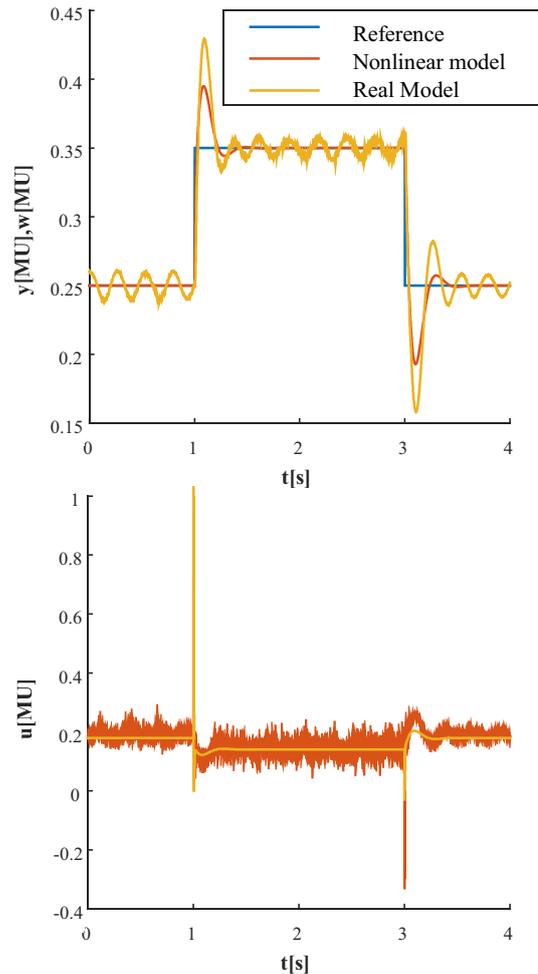


Figure.13: Control experiment

Validation was also performed by control experiments on the system and then the same experiments are repeated on the model. Compared values are output values and input action values as presented in Figure 13. It can be concluded that even from the perspective of validation is created model sufficiently adequate, because both the output and even the control input signal have the same behaviour as the output and control input signal obtained from the real system.

CONCLUSION

The CE152 Magnetic levitation system was investigated and its first principle model was derived. This model was created in the MATLAB/Simulink environment. Obtained model was made more precise by incorporating data from several real-time experiments.

Validation and verification proved a good correspondence of the nonlinear Simulink model and the real time plant.

The model will be further improved and used or control design of a model predictive controller for the magnetic levitation system.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014).

REFERENCES

Bobál, V.; J. Böhm; J. Fessl and J. Macháček. 2005. *Digital Self-tuning Controllers: Algorithms, Implementation and Applications*. Springer - Verlag London Ltd., London.

- Liu, G. P. 2001. *Nonlinear identification and control – A neural network Approach*. Springer - Verlag London Ltd., London.
- Ljung, L. 1999. *System identification: theory for the user*. Upper Saddle River, N.J.: Prentice Hall PTR.
- Himmelblau, D. M. and J. B. Riggs. 2004. *Basic principles and calculations in chemical engineering*, Upper Saddle River, N.J.: Prentice Hall.
- Humusoft. 1996. *CE 152 Magnetic levitation model educational manual*
- Tan, K. C. and Y. Li. 2002. “Grey-box model identification via evolutionary computing.” *Control Engineering Practice*, 10, 673–684.

AUTHOR BIOGRAPHIES



Petr Chalupa was born in Zlin in 1976 and graduated from 1999 from Brno University of Technology and received the Ph.D. degree in Technical cybernetics from Tomas Bata University in Zlin in 2003.

He worked as a programmer and designer of an attendance system and as a developer of a wireless alarm system. He was a researcher at the Centre of Applied Cybernetics. Currently he works a researcher at the Faculty of Applied Informatics, Tomas Bata University in Zlin as a member of CEBIA-Tech team. His research interests are modelling and adaptive and predictive control of real-time systems.

Martin Malý graduated from Tomas Bata University in Zlin, Faculty of Applied Informatics in 2015. Nowadays he works as an engineer in TES Vsetin, Czech Republic



Jakub Novak was born in 1976 and received the Ph.D. degree in chemical and process engineering from Tomas Bata University in Zlin in 2007.

He is a researcher at the Faculty of Applied Informatics, Tomas Bata University in Zlin under a CEBIA-Tech project. His research interests are modeling and predictive control of the nonlinear systems.

LINEAR PREDICTIVE CONTROL OF NONLINEAR TIME-DELAYED MODEL OF LIQUID-LIQUID STIRRED HEAT EXCHANGER

Radek Holíš, Vladimír Bobál
Department of Process Control
Faculty of Applied Informatics, Tomas Bata University in Zlin
NadStráněmi 4511, Zlin 76005, Czech Republic
E-mail: rholis@fai.utb.cz

KEYWORDS

Model Predictive Control, MPC, Time Delay, Heat Exchanger, Parameter Estimation, Nonlinear System, Nonlinear Model, Control Simulation

ABSTRACT

Many nonlinear processes in industry exhibit time delay in their dynamic behavior. Time delay is mainly caused by the time required to transport energy, information or mass, but it can be caused by processing time as well. There are also many cases when compensation of measurable disturbance is required. In case, when the nonlinearity of system is not significant, it can be controlled by linear control method, but accurate system identification in the operating area should be performed. Moreover the suitable control algorithm, which is able to handle these requirements, should be used. In light of these facts, the goal of this paper is to design predictive algorithm for control of nonlinear time-delayed systems with the possibility of measurable disturbance compensation. Basically this paper deals with the essential principles of model predictive control (MPC), design process of the predictive controller, calculation of control law and identification of control process. Identification of control process is also an important part of MPC. This paper describes simulation and verification of designed regulator which was verified in MATLAB/SIMULINK as well.

INTRODUCTION

Heat exchangers are an essential part of many technologies in energy and chemical industry, polymer manufacturing, petroleum refineries, and many others and they are typical examples of nonlinear system. Nonlinear processes with time delay are difficult to control using standard feedback controllers. When a high performance of the control process is desired or the relative time delay is very large, we can choose another approach. One of the most known methods for control of processes with time delay is Smith predictor, however this predictor does not allow to control nonlinear processes. Another negative aspect is effect of measurable disturbance on the process output which is a very important issue that needs to be analyzed and considered in control problems where it is possible to be measured. Disturbances drive the system away from its desired operating point and require more sophisticated

control strategies to minimize their influences. The most known solution to eliminate this problem is classical feedforward control of measurable disturbances. These types of compensators provide a possibility to take control actions before the disturbance affects the process output. On the other hand, for simple processes, where effect of disturbance is not too significant, using of classical feedforward control is sufficient. For more complex processes which are slightly nonlinear, exhibit time delay and with possibility of measuring disturbance, Model Predictive Control (MPC) strategy can be used. This paper deals with the use of MPC for slightly nonlinear processes with time delay with possibility of measuring disturbance compensation. Strategy of MPC presents a series of advantages over other methods. The MPC can be used to control a great variety of processes, ranging from those with relatively simple dynamics to other more complex ones, including systems with long time-delay, unstable ones or non-minimum phase. The multivariable case can easily be dealt with. The additional advantage is that extension to the treatment of constraints is conceptually simple and these can be systematically included during the design process. This approach of control is a totally open methodology based on certain basic principles that is allowed for future extensions.

The MPC was mainly deployed on slower processes. It was caused by the large computational complexity of control algorithms. Over the years trends have expanded towards modifications of MPC, which can control very fast processes (e.g. explicit MPC can be used). In practice, MPC has proven to be a reasonable strategy for industrial control, and several reports indicate that it is the most used advanced control technology in industry (Camacho and Normey-Rico 2007; Rossiter 2003).

Extended version of Generalized Predictive Control (GPC) algorithm is used for design of predictive controller in this paper.

The paper is organized in the following way. The general principle of the MPC is described in the first section. The next section is devoted to explain the extended GPC algorithm. The GPC cost function and control law are introduced in following section. The identification of model of experimental laboratory liquid-liquid stirred heat exchanger is described in the next section. Verification of control method is shown next. Results evaluation is described in following section and the last section concludes the paper.

MODEL PREDICTIVE CONTROL PRINCIPLES

The term MPC does not describe a specific control strategy but a very extensive range of control methods that make explicit use of a model of the process to obtain the control signal by minimizing an objective function. The essential ideas of the predictive control family are (Camacho and Normey-Rico 2007; Rossiter 2003):

- Explicit use of a model to predict the process output at future time instants (horizon).
- The trajectory of the reference signal is known for the time horizon of the prediction.
- Calculation of a control sequence minimizing an objective function, mostly quadratic.
- Only the first computed system input value is used for control and the calculation is repeated in the next sampling period (Camacho and Normey-Rico 2007).

Various discrete models can be used to represent the process behavior which is the main difference between the MPC methods. For example step response, transfer function, impulse response and other discrete models can be used.

Fig. 1 and Fig. 2 represent principle of MPC.

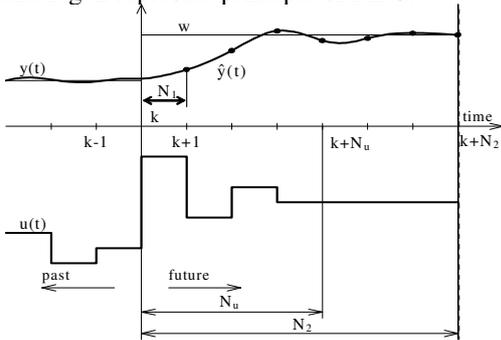


Figure 1 : The MPC Main Idea

- The future outputs for the prediction horizon N are predicted at each instant t using the process model. These predicted outputs $\hat{y}(k+j)$ for $j=1\dots N$ depend on the known values up to instant k (past inputs and outputs) and on the future control signals $u(k+j)$, $j=0\dots N-1$, which are to be sent to the system and to be calculated, but only the first calculated one will be implemented.
- The vector of future control signals is calculated by optimizing a determined criterion in order to keep the process as close as possible to the reference trajectory $w(k+j)$. This criterion usually takes the form of quadratic function of the errors between the predicted output signal and the predicted reference trajectory. The control effort is included in the objective function in most cases.
- The control signal $u(k)$ is sent to the process while the next control signals calculated are neglected, because at the next sampling instant $y(k+1)$ will already be known, and step 1 will be repeated with this new value and all the

sequences will be brought up to date. Therefore the next control signal $u(k+1)$ is calculated using the receding horizon concept (Bars et al. 2011; Bobál 2008; Camacho and Normey-Rico 2007; Schwarz et al. 2012).

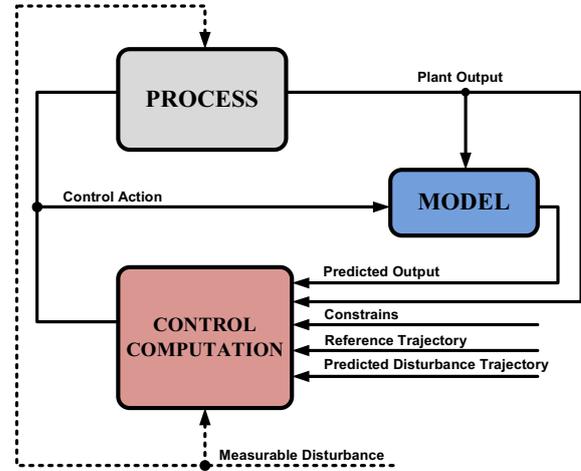


Figure 2 : Extended Structure of MPC

EXTENDED GENERALIZED PREDICTIVE CONTROL ALGORITHM

Principle scheme of extended MPC algorithm is depicted in Fig. 2, where dashed line represents measurable disturbance.

The GPC minimizes a cost function that can be rewritten as

$$J(N_1, N_2, N_u) = \sum_{i=N_1}^{N_2} \delta(i) [\hat{y}(k+i) - w(k+i)]^2 + \sum_{i=1}^{N_u} \lambda(i) [\Delta u(k+i-1)]^2 \quad (1)$$

where $\hat{y}(k+i)$ is an optimum prediction of the system output, N_1 and N_2 are the minimum and maximum costing horizons, N_u is the control horizon, $\delta(i)$ and $\lambda(i)$ are weighting coefficients and $w(k+i)$ is a vector of future reference sequence. The objective of predictive control is to compute the future incremental control sequence $\Delta u(k)$, $\Delta u(k+1)$, ... This is accomplished by minimizing of cost function J . To minimize J , the predictions $\hat{y}(k+i)$ are first expressed as a function of the past data and the future control actions $\Delta u(k+i-1)$. Then, J can be considered as a function of the future control sequence and minimization is accomplished in order to obtain the optimal value. The weighting factors and horizons are the tuning parameters (Camacho and Normey-Rico 2007).

Without loss of generality and because of the time delay characteristics of the process, horizons N_1 and N_2 are computed as $N_1 = d+1$ and $N_2 = N_u + d$.

The extended mathematical model used by GPC to compute the predictions is a modified CARIMA model. It is a typical CARIMA model extended by the vector $v(k)$ which represents measurable disturbance.

$$A(z^{-1})y(k) = z^{-d}B(z^{-1})u(k-1) + z^{-dv}D(z^{-1})v(k) + \frac{C(z^{-1})}{\Delta}e_s(k) \quad (2)$$

where $\Delta = 1 - z^{-1}$, d and dv is number of steps of the time delay and $e_s(k)$ is the white noise. The polynomial $D(z^{-1})$ represents character of disturbance and the polynomial $C(z^{-1})$ describes character of the noise. This character is difficult to determine, therefore polynomial $C(z^{-1})$ was chosen to be equal to one. Vector $v(k)$ is discrete set of measurable disturbance (Camacho and Bordons 2004; Fikar and Mikleš 2008). After application of following equation and multiplication equation (2) by Δ

$$\tilde{A}(z^{-1}) = (1 - z^{-1})A(z^{-1}) = 1 - \tilde{a}_1 z^{-1} \dots - \tilde{a}_{na+1} z^{-na-1} = 1 - (1 - a_1)z^{-1} - (a_1 - a_2)z^{-2} \dots - a_{na+1} z^{-na-1} \quad (3)$$

model should be represented in following way, where output should be predicted as

$$\hat{y}(k+1) = \sum_{i=1}^{na+1} \tilde{a}_i y(k+1-i) + \sum_{i=1}^{nb} b_i \Delta u(k-d-i) + \sum_{i=1}^{nd} d_i \Delta v(k+1-dv-i) \quad (4)$$

where na , nb , nc and nd are degrees of polynomials $A(z^{-1})$, $B(z^{-1})$, $C(z^{-1})$ and $D(z^{-1})$. The white noise $e_s(k)$ and its future values are considered to be equal to zero for the prediction of the future output values.

In case where time delay is present, following equation should be used when equation (4) is applied recursively for $i = 1, 2, \dots, N_u$

$$\begin{bmatrix} \hat{y}(k+d+1) \\ \hat{y}(k+d+2) \\ \vdots \\ \hat{y}(k+d+N_u) \end{bmatrix} = \mathbf{G} \begin{bmatrix} \Delta u(k) \\ \Delta u(k+1) \\ \vdots \\ \Delta u(k+N_u-1) \end{bmatrix} + \mathbf{H} \begin{bmatrix} \Delta u(k-1) \\ \Delta u(k-2) \\ \vdots \\ \Delta u(k-nb) \end{bmatrix} + \mathbf{S} \begin{bmatrix} \hat{y}(k+d) \\ \hat{y}(k+d-1) \\ \vdots \\ \hat{y}(k+d-na) \end{bmatrix} + \mathbf{H}_{v1} \begin{bmatrix} \Delta v(k-1) \\ \Delta v(k-2) \\ \vdots \\ \Delta v(k-nd+1) \end{bmatrix} + \mathbf{H}_{v2} \begin{bmatrix} \Delta v(k) \\ \Delta v(k+1) \\ \vdots \\ \Delta v(k+N_u-1) \end{bmatrix} \quad (5)$$

Equation (5) should be written as

$$\hat{\mathbf{y}} = \mathbf{G}\mathbf{u} + \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 + \mathbf{H}_{v1}\mathbf{v}_1 + \mathbf{H}_{v2}\mathbf{v}_2 \quad (6)$$

Matrices \mathbf{G} , \mathbf{H} and \mathbf{S} are constant matrices of dimensions $N_u \times N_u$, $N_u \times nb$ and $N_u \times (na+1)$. Matrices \mathbf{H}_{v1} and \mathbf{H}_{v2} are of dimensions $N_u \times (nd-1)$ and $N_u \times N_u$. Matrix \mathbf{H}_{v1} can be used only in case when degree of polynomial $D(z^{-1})$ is equal to 2 or higher.

Following equation corresponds to the free response of the system that is the output that would be obtained if the control and disturbance signals were kept constant.

$$\mathbf{f} = \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 + \mathbf{H}_{v1}\mathbf{v}_1 \quad (7)$$

Forced response of system is represented by next equation

$$\mathbf{f}_r = \mathbf{G}\mathbf{u} + \mathbf{H}_{v2}\mathbf{v}_2 \quad (8)$$

Based on equation mentioned above (7) and (8), overall response of system is computed as sum of free and forced response.

$$\hat{\mathbf{y}} = \mathbf{f} + \mathbf{f}_r \quad (9)$$

COST FUNCTION AND CONTROL LAW

If $\hat{\mathbf{y}}$ is introduced in equation (1), it is evident that J is a cost function of \mathbf{y}_1 , \mathbf{u} and \mathbf{u}_1 . Function should be written as

$$J = (\mathbf{G}\mathbf{u} + \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 - \mathbf{w})^T \mathbf{Q}_\delta (\mathbf{G}\mathbf{u} + \mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 - \mathbf{w}) + \mathbf{u}^T \mathbf{Q}_\lambda \mathbf{u} \quad (10)$$

where \mathbf{Q}_δ and \mathbf{Q}_λ are the diagonal weighting matrices of size $N_u \times N_u$ with elements $\delta(j)$ and $\lambda(j)$, respectively. Although, in practice, the most common choice is to set $\delta(j)$ and $\lambda(j)$ constants on the horizon.

In practice, the values of these weighting factors must be normalized in order to obtain a correct weighting of the different errors and controller outputs.

After some manipulations J is written as

$$J = \mathbf{u}^T (\mathbf{Q}_\lambda + \mathbf{G}^T \mathbf{Q}_\delta \mathbf{G}) \mathbf{u} + 2(\mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 - \mathbf{w})^T \mathbf{Q}_\delta \mathbf{G} \mathbf{u} + (\mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 - \mathbf{w})^T \mathbf{Q}_\delta (\mathbf{H}\mathbf{u}_1 + \mathbf{S}\mathbf{y}_1 - \mathbf{w}) \quad (11)$$

Minimizing J with respect to \mathbf{u} , it means $\frac{\partial J}{\partial \mathbf{u}} = 0$,

leads to

$$\mathbf{M}\mathbf{u} = \mathbf{P}_0\mathbf{y}_1 + \mathbf{P}_1\mathbf{u}_1 + \mathbf{P}_2\mathbf{w} \quad (12)$$

where $\mathbf{M} = \mathbf{G}^T \mathbf{Q}_\delta \mathbf{G} + \mathbf{Q}_\lambda$ is of dimension $N_u \times N_u$, $\mathbf{P}_0 = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{S}$ of dimension $N_u \times (na+1)$, $\mathbf{P}_1 = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H}$ of dimension $N_u \times nb$ and $\mathbf{P}_2 = \mathbf{G}^T \mathbf{Q}_\delta$ of dimension $N_u \times N_u$, therefore

$$\mathbf{M} \begin{bmatrix} \Delta u(k) \\ \Delta u(k+1) \\ \vdots \\ \Delta u(k+N_u-1) \end{bmatrix} = \mathbf{P}_0 \begin{bmatrix} \hat{y}(k+d) \\ \hat{y}(k+d-1) \\ \vdots \\ \hat{y}(k+d-na) \end{bmatrix} + \mathbf{P}_1 \begin{bmatrix} \Delta u(k-1) \\ \Delta u(k-2) \\ \vdots \\ \Delta u(k-nb) \end{bmatrix} + \mathbf{P}_2 \begin{bmatrix} w(k+d+1) \\ w(k+d+2) \\ \vdots \\ w(k+d+N_u) \end{bmatrix} \quad (13)$$

In a receding horizon algorithm only the actual value $\Delta u(k)$ is computed, so if \mathbf{m} is the first row of matrix \mathbf{M}^{-1} , then $\Delta u(k)$ is given by

$$\Delta u(k) = \mathbf{m}\mathbf{P}_0\mathbf{y}_1 + \mathbf{m}\mathbf{P}_1\mathbf{u}_1 + \mathbf{m}\mathbf{P}_2\mathbf{w} \quad (14)$$

When compensation of measurable disturbance is required, $\Delta u(k)$ is given by extended form of control law

$$\Delta u(k) = \mathbf{m}\mathbf{P}_0\mathbf{y}_1 + \mathbf{m}\mathbf{P}_1\mathbf{u}_1 + \mathbf{m}\mathbf{P}_2\mathbf{w} + \mathbf{m}\mathbf{P}_{v1}\mathbf{v}_1 + \mathbf{m}\mathbf{P}_{v2}\mathbf{v}_2 \quad (15)$$

where $\mathbf{P}_{V1} = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H}_{V1}$ is of dimension $N_u \times (nd - 1)$ and $\mathbf{P}_{V2} = -\mathbf{G}^T \mathbf{Q}_\delta \mathbf{H}_{V2}$ of dimension $N_u \times N_u$. After introducing vectors \mathbf{y}_1 , \mathbf{u}_1 , \mathbf{w} , \mathbf{v}_1 and \mathbf{v}_2 , final control law is defined as

$$\begin{aligned} \Delta u(k) = & \mathbf{mP}_0 \begin{bmatrix} \hat{y}(k+d) \\ \hat{y}(k+d-1) \\ \vdots \\ \hat{y}(k+d-na) \end{bmatrix} + \mathbf{mP}_1 \begin{bmatrix} \Delta u(k-1) \\ \Delta u(k-2) \\ \vdots \\ \Delta u(k-nb) \end{bmatrix} + \\ & + \mathbf{mP}_2 \begin{bmatrix} w(k+d+1) \\ w(k+d+2) \\ \vdots \\ w(k+d+N_u) \end{bmatrix} + \mathbf{mP}_{V1} \begin{bmatrix} \Delta v(k-1) \\ \Delta v(k-2) \\ \vdots \\ \Delta v(k-nd+1) \end{bmatrix} + \\ & + \mathbf{mP}_{V2} \begin{bmatrix} \Delta v(k) \\ \Delta v(k+1) \\ \vdots \\ \Delta v(k+N_u-1) \end{bmatrix} \end{aligned} \quad (16)$$

\mathbf{H}_{V1} and \mathbf{H}_{V2} are matrices including the coefficients of the system step response to the disturbance.

Future values of disturbance can be known in some cases, when disturbance is related to the process load. In other cases, it can be predicted by using means, trends or other information. If this is the case, the term corresponding to future deterministic disturbance can be computed (Pawlowska et al. 2012).

It is evident that matrices \mathbf{H}_{V1} and \mathbf{H}_{V2} are dependent on the relative difference between number of steps of time delay of input-output and disturbance-output which is defined as

$$\rho = d - dv \quad (17)$$

This leads to three different structures for matrices \mathbf{H}_{V1} and \mathbf{H}_{V2} based on the sign of ρ :

- $\rho < 0$

$$\mathbf{H}_{VX} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 \\ h_1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & 0 & 0 & 0 & 0 \\ h_{N_u+\rho} & \dots & h_1 & 0 & 0 & 0 \end{bmatrix} \left. \vphantom{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 \\ h_1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & 0 & 0 & 0 & 0 \\ h_{N_u+\rho} & \dots & h_1 & 0 & 0 & 0 \end{bmatrix}} \right\} |\rho| \quad (18)$$

- $\rho = 0$

$$\mathbf{H}_{VX} = \begin{bmatrix} h_1 & 0 & 0 & 0 & 0 & 0 \\ h_2 & \ddots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & h_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ h_{N_u} & h_{N_u-1} & \dots & \dots & h_2 & h_1 \end{bmatrix} \quad (19)$$

- $\rho = 0$

$$\mathbf{H}_{VX} = \begin{bmatrix} h_{\rho+1} & h_\rho & \dots & h_1 & 0 & 0 \\ h_{\rho+2} & h_{\rho+1} & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & h_1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & h_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N_u+\rho} & h_{N_u+\rho-1} & \dots & \dots & h_\rho & h_{\rho+1} \end{bmatrix} \left. \vphantom{\begin{bmatrix} h_{\rho+1} & h_\rho & \dots & h_1 & 0 & 0 \\ h_{\rho+2} & h_{\rho+1} & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & h_1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & h_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{N_u+\rho} & h_{N_u+\rho-1} & \dots & \dots & h_\rho & h_{\rho+1} \end{bmatrix}} \right\} \rho \quad (20)$$

Where h_i are the coefficients of \mathbf{H}_{V1} and \mathbf{H}_{V2} matrices obtained from the delay free disturbance response shifted in accordance with ρ (Pawlowska et al. 2012).

IDENTIFICATION

Since MPC is based on an internal model of the system, a sufficient model is necessary therefore system identification is an important factor in MPC strategy.

Identification of control processes can be divided into two groups that are used most often. The first group is one-time (offline) method and the second is ongoing (online) methods. Online identification methods can be used for self-tuning controllers (STC).

Both types of identification can be used for estimation of the model parameters of the system. Online identification, offline identification and comparison with estimation by MATLAB function *fminsearch* are shown in following subsections.

Offline Identification Methods

The most used method for the identification of the parameters of discrete transfer function models is the least squares estimator (LSE) based on the idea of linear regression. These identification algorithms can be carried out in an online manner as well.

The least squares estimator is defined as the vector $\hat{\Theta}$ that minimizes the quadratic error

$$\hat{\Theta} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{y} \quad (21)$$

Notice that $\hat{\Theta}$ is a vector of estimated model parameters of dimension $2n$, \mathbf{F} is a matrix of dimension $N - n - d \times 2n$, \mathbf{y} is a data vector of dimension $N - n - d$, where N is a number of measured data, n is an order of system and d is a number of steps of time delay (Camacho and Normey-Rico 2007).

MATLAB function *fminsearch* can find the minimum of a scalar function of several variables as well, starting at an initial estimate. This function uses the simplex search method for finding the minimum of a function.

$$x = \text{fminsearch}(\text{fun}, x_0, \text{options}) \quad (22)$$

Online Identification Methods

One of the advantages of the identification procedure based on the LSM is that it can be used recursively because the parameter vector estimated at time t can be computed as a function of the parameter vector estimated at $k - 1$. The recursive least squares method (RLSM) is the most known method. This method uses ARX model (Bobál et al. 2012).

$$y(k) = \Theta^T(k) \Phi(k) + e_s(k) \quad (23)$$

where Θ is a vector of model parameters

$$\Theta^T(k) = [a_1 \ a_2 \dots a_n \ b_1 \ b_2 \dots b_n] \quad (24)$$

and Φ is a regression vector

$$\Phi^T(k) = [-y(k-1) \ -y(k-2) \dots -y(k-n) \ u(k-d-1) \ u(k-d-2) \dots u(k-d-n)] \quad (25)$$

Identification of Model of Liquid-Liquid Stirred Heat Exchanger

Identification was performed on the SIMULINK model of stirred heat exchanger and the estimated mathematical model was used for verification purposes.

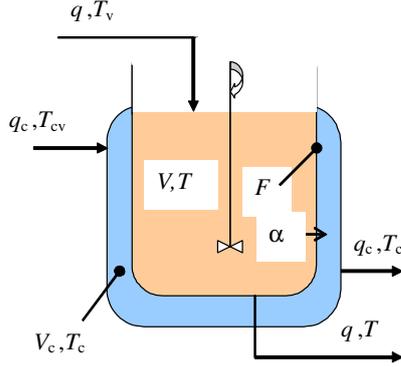


Figure 3 : Model of Stirred Flow Heat Exchanger

Model of this heat exchanger can be described by two differential equations of first order:

$$q \rho c_p T_v = q \rho c_p T + F \alpha (T - T_c) + V \rho c_p \frac{dT}{dt} \quad (26)$$

$$q_c \rho_c c_{pc} T_{cv} + F \alpha (T - T_c) = q_c \rho_c c_{pc} T_c + V_c \rho_c c_{pc} \frac{dT_c}{dt} \quad (27)$$

with initial conditions:

$$T(0) = T^s, \ T_c(0) = T_c^s \quad (28)$$

and boundary conditions:

$$0.02 \leq q_c \leq 0.6 \quad (29)$$

where t stands for the time, T for temperatures, q for flow of fluids, ρ for densities, c_p for specific heat capacities, α for heat transfer coefficients, V for volumes and F for heat exchanger area.

For the control purposes, the output temperature of the refrigerated fluid $T(t)$ 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.

Parameters of heat exchanger were chosen as is shown in the following table.

Table 1 : Parameters of Heat Exchanger

Parameter	Value
$V = 2.65 \text{ m}^3$	$q = 0.2 \text{ m}^3 \text{ min}^{-1}$
$V_c = 0.63 \text{ m}^3$	$q_c = 0.4 \text{ m}^3 \text{ min}^{-1}$
$\rho = 985 \text{ kg m}^{-3}$	$c_p = 4.05 \text{ kJ kg}^{-1} \text{ K}^{-1}$
$\rho_c = 998 \text{ kg m}^{-3}$	$c_{pc} = 4.18 \text{ kJ kg}^{-1} \text{ K}^{-1}$
$F = 8.8 \text{ m}^2$	$\alpha = 58 \text{ kJ m}^{-2} \text{ min}^{-1} \text{ K}^{-1}$
$T_v = 370.0 \text{ K}$	$T_c = 293.0 \text{ K}$

Steady-State Characteristic of Model of Liquid-Liquid Stirred Heat Exchanger

The dependence of the refrigerated fluid output temperature on the coolant flow in the steady-state is in Figure. 4.

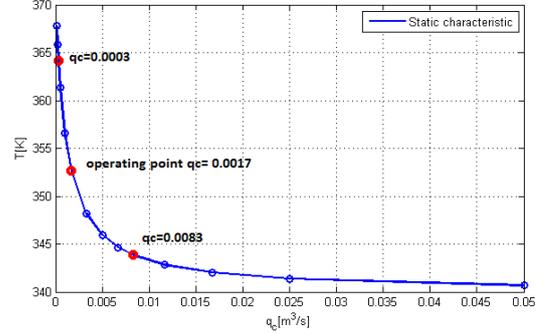


Figure 4 : Static Characteristic of Heat Exchanger

In subsequent control simulations, the operating interval for q_c has been determined as $0.0003 \leq q_c \leq 0.0083$, where the static characteristic is only slightly nonlinear as is shown in Figure 5.

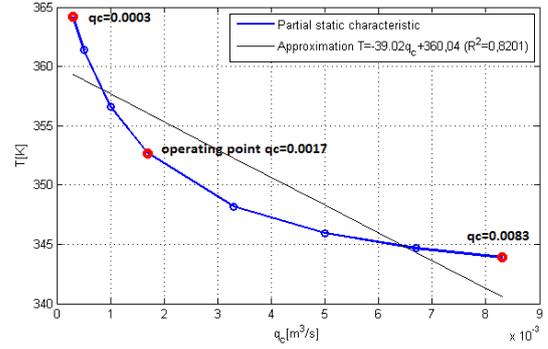


Figure 5 : Static Characteristic for Operating Interval

Choice of Input Sequence for Model of Liquid-Liquid Stirred Heat Exchanger Identification

Pseudorandom Binary Signal (PRBS), Random Binary Signal (RBS) and Random Gaussian Signal (RGS) were used for identification of laboratory model, see Table 6-8. Input signals were generated by MATLAB function *idinput*.

$$u = \text{idinput}(N, \text{type}, \text{band}, \text{levels}) \quad (30)$$

Function *idinput* generates input signals which are used for identification purposes, where u is returned as matrix or column vector. Parameter N determines the number of generated input data and parameter *type* defines the type of input signal to be generated.

Sum of squares subtraction of estimated outputs and measured data was used for analysis of quality of identified models.

$$S_y = \frac{1}{N} \sum_{k=a}^b [\hat{y}(k) - y(k)]^2 \quad (31)$$

Where N represents number of measured data, $y(k)$ is measured output value and $\hat{y}(k)$ is estimated output value.

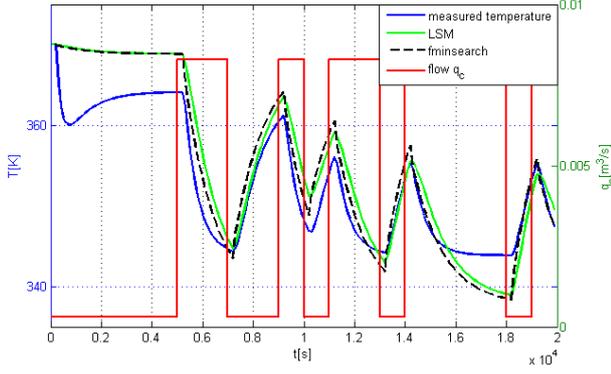


Figure 6 : Identification by PRBS Input Signal

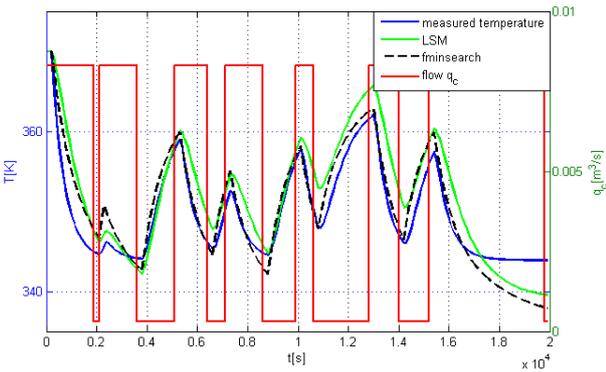


Figure 7 : Identification by RBS Input Signal

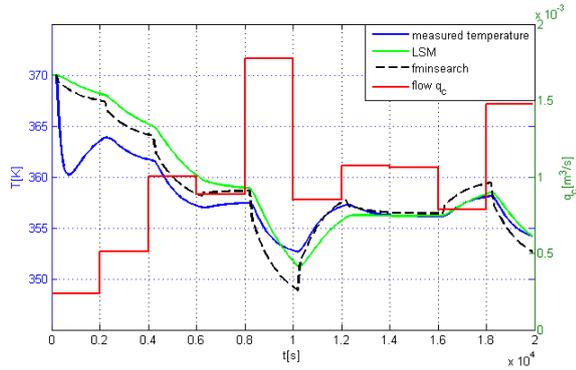


Figure 8 : Identification by RGS Input Signal

Following table shows results of laboratory model identification.

Table 2 : Evaluation of Identification Quality

Method	Parameter	PRBS	RBS	RGS	
LSM	Coeffs.	a_1	-1.6166	-1.6059	-1.5631
		a_2	0.6416	0.6319	0.5839
		b_1	-149.9939	-126.3404	-276.0283
		b_2	54.6479	29.0058	6.4397
	Quality S_v	18.9143	13.5954	8.1969	
RLSM	Coeffs.	a_1	-1.3710	-1.2834	-1.5538
		a_2	0.4183	0.3368	0.5779
		b_1	-144.4788	-157.9094	-202.3707
		b_2	-40.0787	-47.1242	-109.0916
	Quality S_v	17.6613	10.7138	7.7424	
fminsearch	Coeffs.	a_1	-0.9825	-1.7038	-0.8247
		a_2	0.0487	0.7121	-0.0858
		b_1	-889.1335	-485.4889	-4823.0477
		b_2	630.3898	450.0139	3678.7293
	Quality S_v	16.1456	6.8179	5.7447	

SIMULATION VERIFICATION

Parameters estimated by RLSM with RBS were chosen for verification purposes since the frequency spectrum of RBS is suitable for these types of systems. Discrete transfer function for exchanger has the following form:

$$G_S(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} z^{-d} = \frac{-157.9094z^{-1} - 47.1242z^{-2}}{1 - 1.2834z^{-1} + 0.3368z^{-2}} z^{-2} \quad (32)$$

Discrete transfer function for disturbance is:

$$G_V(z^{-1}) = \frac{D(z^{-1})}{A(z^{-1})} z^{-dv} = \frac{-78.9547z^{-1} - 23.5621z^{-2}}{1 - 1.2834z^{-1} + 0.3368z^{-2}} z^{-2} \quad (33)$$

where $T_S = 100s$, $d = 2$, $dv = 2$, $\delta = 1$, $\lambda = 20$ and $N_u = 10$.

Figures below show regulation processes for various types of disturbance compensation (predictive algorithm without compensation of measurable disturbance (A), with compensation (B) and the case when vector of measurable disturbance is known in time (C)).

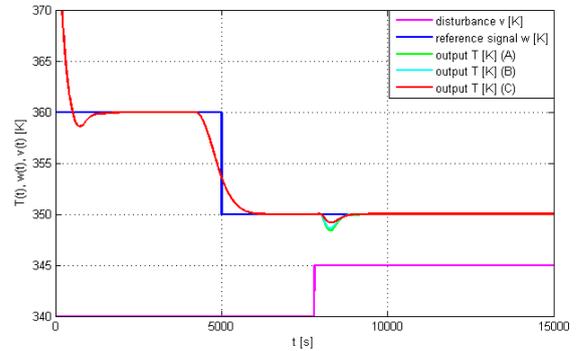


Figure 9 : Regulation Processes (A, B and C)

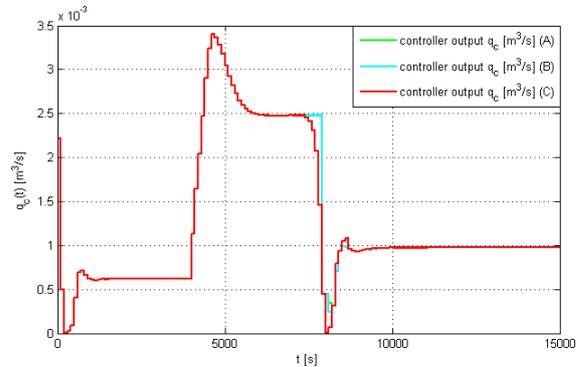


Figure 10 : Controller Outputs (A, B and C)

ANALYSIS OF RESULTS

The quadratic criterion of the measured error, the quadratic criterion of the system output increments and absolute values of these signals were chosen for a control quality analysis.

$$S_{e2} = \frac{1}{N} \sum_{k=a}^b [e(k)]^2, \quad S_{ea} = \frac{1}{N} \sum_{k=a}^b |e(k)| \quad (34)$$

$$S_{du2} = \frac{1}{N} \sum_{k=a}^b [\Delta u(k)]^2, \quad S_{dua} = \frac{1}{N} \sum_{k=a}^b |\Delta u(k)|$$

Evaluation of control quality for various types of disturbance compensation is shown in the following table.

Table 3 :Evaluation for Simulation of Heat Exchanger

	Overshoot	S_{e2}	S_{ea}	$S_{du2} \cdot 10^{-8}$	$S_{dua} \cdot 10^{-5}$
Without com. (A)	2.11	13.22	2.59	6.02	4.00
With comp. (B)	1.52	13.14	2.56	6.02	4.00
Known dist. (C)	1.27	13.02	2.49	7.79	4.56

CONCLUSION

This paper dealt with an extended GPC design procedure to improve the measurable disturbance compensation. The extended GPC control law with implicit disturbance compensation is interpreted as classical feedback plus feedforward control scheme that is described by theoretical analysis. Functionality of designed algorithm was simulation verified on a mathematical model of liquid-liquid stirred heat exchanger.

In case, where impact of different types of disturbance compensation was verified, there is shown that integration of disturbance compensation to basic GPC algorithm improves regulation processes and control quality as well. Without compensation of measurable disturbance, i.e. standard version of GPC, overshoot of output signal is 2.11 K. When compensation of disturbance is used, but vector of disturbance is unknown (the case when only disturbance is measured), overshoot is reduced to 1.52 K. In situation, when vector of disturbance is known, compensation is more substantial (overshoot 1.27 K) and can have positive effect when it is measured and predicted simultaneously, e.g. sunrise or sunset and effect of sun can be predicted based on part of season; particular substance is inserted into chemical reactor at a certain time; vehicles can predict and adapt driving style based on type of turn on road and many other cases.

The simulation results demonstrate improved usability of extended GPC algorithm and results also showed that this algorithm is able to control slightly nonlinear processes.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Czech Republic under grant IGA/FAI/2016/006.

REFERENCES

- Bars R.; R. Haber and U. Schmitz. 2011. *Predictive control in process engineering: From the basics to the applications*. Weinheim: Willey-VCH Verlag.
- Bitmead, R.R.; M. Gevers and V. Hertz. 1990. *Adaptive optimal control. The thinking man's GPC*, Prentice Hall, Englewood Cliffs, New Jersey.
- Bobál, V. 2008, *Adaptive and predictive control*. vol. 1. Zlín, Tomas Bata University in Zlín (in Czech).
- Bobál, V.; P. Chalupa; M. Kubalčík and P. Dostál. 2012. "Identification and self-tuning control of time-delay systems", WSEAS Transactions on Systems, vol. 11, pp. 596-608
- Camacho E.F. and C. Bordons. 2004. *Model predictive control*, Springer Verlag, London.

Camacho E.F. and J.E. Normey-Rico. 2007. *Control of dead-time processes*, Springer-Verlag, London.

Clarke D.W.; C. Mohtadi and P.S. Tuffs. 1987. "Generalized predictive control, part I: the basic algorithm", *Automatica*, vol. 23, pp. 137-148.

Clarke D.W.; C. Mohtadi and P.S. Tuffs. 1987. "Generalized predictive control, part II: extensions and interpretations", *Automatica*, vol. 23, pp. 149-160.

Fikar M. and J. Míkleš. 2008. *Process modelling, optimisation and control*, Springer-Verlag, Berlin.

Moudgalya K.M. 2007. *Digital control*. Chichester: John Wiley.

Pawlowska A.; Guzmána J. L.; Normey-Rico J. E. and M. Berenguela. 2012. "Improving feedforward disturbance compensation capabilities in Generalized Predictive Control". *Journal of Process Control*, vol. 22, pp. 527-539.

Holiš R. and V. Bobál. *Possible approaches of disturbance compensation of time-delayed systems using predictive control*. 29th European Conference on Modelling and Simulation (ECMS 2015), 305-311.

Rossiter J.A. 2003. *Model based predictive control: a practical approach*, CRC Press.

Schwarz M. H.; Cox C. S. and J. Börcsök. 2010. "A Filtered Tuning Method for a GPC Controller". In: University of Kassel, Germany, pp. 180-185.

AUTHOR BIOGRAPHIES



RADEK HOLÍŠ studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master's degree in Automatic Control and Informatics in 2014. 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 focuses on modeling and simulation of discrete technological processes, adaptive control and model predictive control. He is currently working in Honeywell HTS-CZ Brno in Aerospace division as Software Design Engineer. His email address is rholis@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.

STATE-SPACE PREDICTIVE CONTROL OF TWO LIQUID TANKS SYSTEM WITH CONSTRAINTS OF PROCESS VARIABLES

Lukáš Rušar, Stanislav Talaš, Adam Krhovják and Vladimír Bobál
 Department of process control
 Faculty of applied informatics, Tomas Bata university in Zlin
 Nad Stráněmi 4511, Zlin 76005, Czech Republic
 E-mail: rusar@fai.utb.cz

KEYWORDS

predictive control, time-delay, constraints of process variables, state-space, two liquid tanks.

ABSTRACT

This paper presents a method called the predictive control used to control a nonlinear process about a selected operating point. The system of the two funnel liquid tanks in series is chosen as an exemplar process. The parameters of the tanks simulate a large industry tanks used in a chemical industry. The state-space CARIMA mathematical model is used for the output values prediction. This paper describes the linearization process of the nonlinear system at the operating point and a possibility of constraints of the process variables. The designed controller is verified on the process without and with a time-delay.

INTRODUCTION

Many processes in the real world are nonlinear, complex and they often include a time-delay. These processes are very difficult to control. The situation is more complicated when some process variables require some sort of a limitation. The basic control methods do not handle with this situation so we need a more advanced method. The predictive control is a great example of the modern control method capable of solving the complex control problem (Bobál 2008).

This method is based on the prediction of the output values on the chosen time horizon. This time horizon should be long enough to cover the step response of the controlled system and the prediction of the output values is based on the mathematical model of the controlled system. The predictive control in this paper uses state-space CARIMA model for multi-input multi-output (MIMO) system (Bars et al. 2011; Wang 2009).

The control signal is obtain by minimization of a cost function. This cost function has usually a quadratic form and it minimize the differences between the reference value and the output value and the control signal increments. We can also take into account the constraints of the process variables in the cost function minimization process. This is done by using a quadratic programming method (Camacho and Bordons 2004; Maciejowski 2002; Rossiter 2003).

However, the state-space CARIMA model is a linear mathematical model. So we have to do one more step to

control the nonlinear system like the chosen system of the two funnel liquid tanks in series. This step is linearization of the nonlinear mathematical model in a selected operating point. The divergence linear model is result of the linearization. It means, that the selected operating point is a new origin state for the controller and the input and the output values are divergence from the equilibrium values (Albertos Pérez and Sala 20014; Hangos et al. 2004).

MATHEMATICAL MODEL OF THE CONTROLLED SYSTEM

The mathematical model of the chosen experimental system of the two liquid tanks system is taken from (Krhovják et al. 2015). This model represents a nonlinear system with two input variables and two output variables. Figure 1 shows a schematic diagram of the controlled process consists of two funnel liquid tanks in series. The first tank is filled by input stream q_{1f} , the second tank is filled by input stream q_{2f} and output stream from the first tank q_1 .

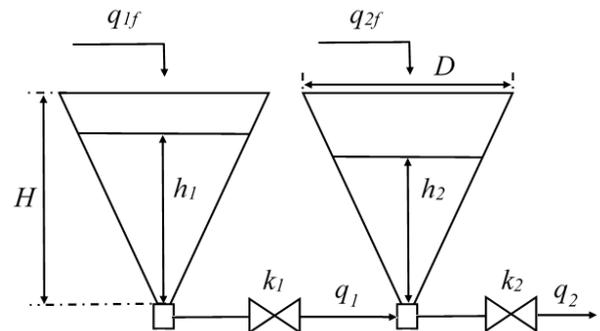


Figure 1 : Schematic diagram of the process

The mathematical model of this process can be obtained from balancing equations of the input and output mass streams. The equation (1) stand for the balancing equation of the input, output and accumulation of the first tank and the equation (2) stand for the balancing equation of the input, output and accumulation of the second tank (Richardson 1989).

$$\pi \frac{D^2}{4H^2} h_1^2 \frac{dh_1}{dt} + q_1 = q_{1f} \quad (1)$$

$$\pi \frac{D^2}{4H^2} h_2^2 \frac{dh_2}{dt} - q_1 + q_2 = q_{2f} \quad (2)$$

In these equations, D is the maximum diameter, H is the total height and h_1 and h_2 are the liquid levels from the bottom of the tanks. The output liquid streams q_1 and q_2 depend on the valve constants k_1, k_2 and the liquid levels as well.

$$q_1 = k_1 \sqrt{|h_1 - h_2|} \quad (3)$$

$$q_2 = k_2 \sqrt{h_2} \quad (4)$$

However, the chosen predictive control method works only with linear mathematical models, so the model of the described process needs to be linearized about an operating point. First of all, the equations (1) and (2) have to be expressed as nonlinear state-space model by selecting the input variables as $u_1 = q_{1f}$ and $u_2 = q_{2f}$ and the output and state variables as $y_1 = x_1 = h_1$ and $y_2 = x_2 = h_2$. This state-space model has form

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ \mathbf{y} &= \mathbf{g}(\mathbf{x}) \end{aligned} \quad (5)$$

where equations of single states and outputs are

$$\begin{aligned} \frac{dx_1}{dt} &= \frac{4H^2}{\pi D^2 x_1^2} (u_1 - k_1 \sqrt{|x_1 - x_2|}) \\ \frac{dx_2}{dt} &= \frac{4H^2}{\pi D^2 x_2^2} (u_2 + k_1 \sqrt{|x_1 - x_2|} - k_2 \sqrt{x_2}) \\ y_1 &= x_1 \\ y_2 &= x_2 \end{aligned} \quad (6)$$

To linearized this nonlinear state-space model, we have to calculate the operating point as equilibrium point of the system. This means, that we are looking for the state, where are no changes of the liquid levels in the tanks and steady input streams to the tanks.

$$\begin{aligned} 0 &= \frac{4H^2}{\pi D^2 x_1^2} (u_1 - k_1 \sqrt{|x_1 - x_2|}) \\ 0 &= \frac{4H^2}{\pi D^2 x_2^2} (u_2 + k_1 \sqrt{|x_1 - x_2|} - k_2 \sqrt{x_2}) \end{aligned} \quad (7)$$

The linearization about the operating point means substitution of the absolute value of the input, output and state variables by its divergence from the steady state.

$$\begin{aligned} \mathbf{x}_\delta(t) &= \mathbf{x}(t) - \bar{\mathbf{x}} \\ \mathbf{u}_\delta(t) &= \mathbf{u}(t) - \bar{\mathbf{u}} \\ \mathbf{y}_\delta(t) &= \mathbf{y}(t) - \bar{\mathbf{y}} \end{aligned} \quad (8)$$

Where $\bar{\mathbf{x}}$ is a vector of the equilibrium state variables, $\bar{\mathbf{u}}$ is a vector of the equilibrium input variables, $\bar{\mathbf{y}}$ is a vector of the equilibrium output variables, $\mathbf{x}_\delta, \mathbf{u}_\delta, \mathbf{y}_\delta$ are divergences from equilibrium values.

The linearized state-space model can be expressed in form

$$\begin{aligned} \dot{\mathbf{x}}_\delta &= \mathbf{A}\mathbf{x}_\delta + \mathbf{B}\mathbf{u}_\delta \\ \mathbf{y}_\delta &= \mathbf{C}\mathbf{x}_\delta \end{aligned} \quad (9)$$

where matrices \mathbf{A}, \mathbf{B} are

$$\begin{aligned} \mathbf{A} &= \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\bar{\mathbf{x}}, \mathbf{u}=\bar{\mathbf{u}}} \\ \mathbf{B} &= \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right|_{\mathbf{x}=\bar{\mathbf{x}}, \mathbf{u}=\bar{\mathbf{u}}} \end{aligned} \quad (10)$$

and \mathbf{C} is an identity matrix (Albertos Pérez and Sala 2004; Hangos et al. 2004).

At this point we transfer this linear state-space model into linear continuous-time input-output model

$$\begin{aligned} \mathbf{A}(s) \mathbf{y}(t) &= \mathbf{B}(s) \mathbf{u}(t) \\ \begin{bmatrix} s + a_{01} & a_{02} \\ a_{03} & s + a_{04} \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} &= \begin{bmatrix} b_{01} & 0 \\ 0 & b_{04} \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \end{aligned} \quad (11)$$

where

$$\begin{aligned} a_{01}(\mathbf{h}) &= \frac{4H^2 k_1 \sqrt{|h_1 - h_2|}}{2D^2 \pi h_1^2 (h_1 - h_2)} \\ a_{02}(\mathbf{h}) &= -\frac{4H^2 k_1 \sqrt{|h_1 - h_2|}}{2D^2 \pi h_1^2 (h_1 - h_2)} \\ a_{03}(\mathbf{h}) &= -\frac{4H^2 k_1 \sqrt{|h_1 - h_2|}}{2D^2 \pi h_2^2 (h_1 - h_2)} \\ a_{04}(\mathbf{h}) &= \frac{4H^2}{2D^2 \pi h_2^2} \left[\frac{k_1 \sqrt{|h_1 - h_2|}}{h_1 - h_2} + \frac{k_2 \sqrt{h_2}}{h_2} \right] \\ b_{01}(\mathbf{h}) &= \frac{4H^2}{D^2 \pi h_1^2}, b_{04}(\mathbf{h}) = \frac{4H^2}{D^2 \pi h_2^2} \end{aligned} \quad (12)$$

This continuous-time model needs to be transferred into a discrete-time input-output model with the sampling period T_0 .

$$\tilde{\mathbf{A}}(z^{-1}) \mathbf{y}(k) = \mathbf{B}(z^{-1}) z^{-d} \Delta \mathbf{u}(k) \quad (13)$$

where the polynomial matrix $\tilde{\mathbf{A}}(z^{-1})$ is

$$\tilde{\mathbf{A}}(z^{-1}) = (1 - z^{-1}) \mathbf{A}(z^{-1}) \quad (14)$$

STATE-SPACE PREDICTIVE CONTROL

The mathematical model used for the prediction of the output values is based on the state-space CARIMA model. It can be expressed as

$$\begin{aligned} \mathbf{x}(k+1) &= \tilde{\mathbf{A}}\mathbf{x}(k) + \mathbf{B}\Delta \mathbf{u}(k-d) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) \end{aligned} \quad (15)$$

where the vector of state variables has form

$$\begin{aligned} \mathbf{x}(k) &= [\mathbf{y}(k), \mathbf{y}(k-1), \dots, \mathbf{y}(k-na), \\ &\Delta \mathbf{u}(k-d-1), \dots, \Delta \mathbf{u}(k-d-nb+1)]^T \end{aligned} \quad (16)$$

and the vectors of the outputs variables and the input control increments are

$$\mathbf{y}(k) = [y_1(k) \quad y_2(k) \quad \dots \quad y_n(k)]^T \quad (17)$$

$$\Delta \mathbf{u}(k-d) = [\Delta u_1(k-d) \quad \dots \quad \Delta u_m(k-d)]^T \quad (18)$$

where n is number of outputs and m is number of inputs (Bars et al. 2011).

The matrices $\tilde{\mathbf{A}}$, \mathbf{B} and \mathbf{C} can be expressed as

$$\tilde{\mathbf{A}} = \begin{bmatrix} -\tilde{\mathbf{A}}_1 & \cdots & -\tilde{\mathbf{A}}_{na} & -\tilde{\mathbf{A}}_{na+1} & \mathbf{B}_2 & \cdots & \mathbf{B}_{nb-1} & \mathbf{B}_{nb} \\ \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}$$

$$\mathbf{B} = [\mathbf{B}_1 \ \mathbf{0} \ \cdots \ \mathbf{0} \ \mathbf{0} \ \mathbf{I} \ \mathbf{0} \ \cdots \ \mathbf{0} \ \mathbf{0}]^T$$

$$\mathbf{C} = [\mathbf{I} \ \mathbf{0} \ \cdots \ \mathbf{0} \ \mathbf{0}] \quad (19)$$

where \mathbf{I} is an identity matrix and $\mathbf{0}$ is a zeros matrix. The matrices $-\tilde{\mathbf{A}}_i$ for $i=1, \dots, na+1$ and \mathbf{B}_j for $j=1, \dots, nb$ consist of the coefficients of the polynoms of the polynomial matrices $\tilde{\mathbf{A}}(z^{-1})$ and $\mathbf{B}(z^{-1})$.

The output values prediction can be calculated recursively using state-space model in equation (15) and it can be expressed in a matrix form

$$\hat{\mathbf{y}} = \mathbf{F}\mathbf{x} + \mathbf{H}_p \Delta \mathbf{u}_p + \mathbf{H}_f \Delta \mathbf{u}_f \quad (20)$$

where $\hat{\mathbf{y}}$ is the vector of the predicted output values, $\Delta \mathbf{u}_p$ is the vector of the past control increments and $\Delta \mathbf{u}_f$ is the vector of the future control increments. These vectors are

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathbf{y}}(k+d+1) \\ \hat{\mathbf{y}}(k+d+2) \\ \vdots \\ \hat{\mathbf{y}}(k+d+N) \end{bmatrix}$$

$$\Delta \mathbf{u}_p = \begin{bmatrix} \Delta \mathbf{u}(k-d) \\ \Delta \mathbf{u}(k-d+1) \\ \vdots \\ \Delta \mathbf{u}(k-1) \end{bmatrix}$$

$$\Delta \mathbf{u}_f = \begin{bmatrix} \Delta \mathbf{u}(k) \\ \Delta \mathbf{u}(k+1) \\ \vdots \\ \Delta \mathbf{u}(k+N) \end{bmatrix} \quad (21)$$

where d is number of steps of the time-delay and N is the prediction time horizon (Camacho and Normey-Rico 2007). This time horizon should be long enough to cover the step response of the controlled system.

This output values prediction can be substituted into the quadratic cost function that the chosen predictive control method minimizes.

$$J = (\mathbf{w} - \hat{\mathbf{y}})^T \mathbf{Q}_\delta (\mathbf{w} - \hat{\mathbf{y}}) + \Delta \mathbf{u}_f^T \mathbf{Q}_\lambda \Delta \mathbf{u}_f \quad (22)$$

where \mathbf{w} is a vector of the future reference values, $\hat{\mathbf{y}}$ is the vector of the predicted outputs values, \mathbf{Q}_λ and \mathbf{Q}_δ are the diagonal weighting matrices. The vector $\Delta \mathbf{u}_f$ is unknown vector of the future control increments (Fikar and Mikleš 2008).

When the constraints of the process variables are not required, the control signal is obtain by minimizing this cost function. But when the constraints of the process variables are necessary, the cost function needs to be modified into form suitable for quadratic programming method

$$J = \frac{1}{2} \mathbf{u}^T \mathbf{H}_c \mathbf{u} + \mathbf{g}^T \mathbf{u} \quad (23)$$

where

$$\mathbf{H}_c = 2(\mathbf{Q}_\lambda + \mathbf{H}_f^T \mathbf{Q}_\delta \mathbf{H}_f)$$

$$\mathbf{g}^T = 2(\mathbf{H}_p \Delta \mathbf{u}_p + \mathbf{F}\mathbf{x} - \mathbf{w})^T \mathbf{Q}_\delta \mathbf{H}_f \quad (24)$$

CONSTRAINTS OF THE PROCESS VARIABLES

The constraints of the process variables mean limitation of the input, output and state values. The most common limitations are

- Limitation of control increments:
 $\Delta \mathbf{u}_{\min} \leq \Delta \mathbf{u}(k) \leq \Delta \mathbf{u}_{\max}$
- Limitation of the absolute control input signal:
 $\mathbf{u}_{\min} \leq \mathbf{u}(k) \leq \mathbf{u}_{\max}$
- Limitation of the output value:
 $\mathbf{y}_{\min} \leq \mathbf{y}(k) \leq \mathbf{y}_{\max}$

All of the constraints need to be expressed as control increments constraints. It means, that it is possible to constrict every variable that depends on the control signal (Camacho and Bordons 2004; Maciejowski 2002). All of these constrains can be joined into one equation for purposes of the quadratic programming

$$\mathbf{A}\mathbf{u} \leq \mathbf{b} \quad (25)$$

where \mathbf{u} is a vector of the future control increments, \mathbf{A} is a corresponding matrix and \mathbf{b} is a vector of the constricted values. Their dimensions depend on desired constraints and number of the inputs and outputs.

The constraints of the control increments can be expressed as

$$\Delta \mathbf{u}(k) \leq \Delta \mathbf{u}_{\max}$$

$$\mathbf{I} \Delta \mathbf{u}_f \leq \Delta \mathbf{u}_{\max} \quad (26)$$

$$\Delta \mathbf{u}(k) \geq \Delta \mathbf{u}_{\min}$$

$$-\Delta \mathbf{u}(k) \leq -\Delta \mathbf{u}_{\min}$$

$$-\mathbf{I} \Delta \mathbf{u}_f \leq -\Delta \mathbf{u}_{\min} \quad (27)$$

The constraints of the absolute value of the control signals can be expressed as

$$\begin{aligned}
\mathbf{u}(k) &\leq \mathbf{u}_{\max} \\
\mathbf{u}(k-1) + \Delta\mathbf{u}(k) &\leq \mathbf{u}_{\max} \\
\Delta\mathbf{u}(k) &\leq \mathbf{u}_{\max} - \mathbf{u}(k-1) \\
\mathbf{T}\Delta\mathbf{u}_f &\leq \mathbf{u}_{\max} - \mathbf{u}_{k-1}
\end{aligned} \tag{28}$$

$$\begin{aligned}
\mathbf{u}(k) &\geq \mathbf{u}_{\min} \\
-\mathbf{u}(k-1) - \Delta\mathbf{u}(k) &\leq -\mathbf{u}_{\min} \\
-\Delta\mathbf{u}(k) &\leq -\mathbf{u}_{\min} + \mathbf{u}(k-1) \\
-\mathbf{T}\Delta\mathbf{u}_f &\leq -\mathbf{u}_{\min} + \mathbf{u}_{k-1}
\end{aligned} \tag{29}$$

The constraints of the output values can be expressed as

$$\begin{aligned}
\mathbf{y}(k) &\leq \mathbf{y}_{\max} \\
\mathbf{H}_f \Delta\mathbf{u}_f + \mathbf{y}_{free} &\leq \mathbf{y}_{\max} \\
\mathbf{H}_f \Delta\mathbf{u}_f &\leq \mathbf{y}_{\max} - \mathbf{y}_{free}
\end{aligned} \tag{30}$$

$$\begin{aligned}
\mathbf{y}(k) &\geq \mathbf{y}_{\min} \\
-\mathbf{H}_f \Delta\mathbf{u}_f - \mathbf{y}_{free} &\leq -\mathbf{y}_{\min} \\
-\mathbf{H}_f \Delta\mathbf{u}_f &\leq -\mathbf{y}_{\min} + \mathbf{y}_{free}
\end{aligned} \tag{31}$$

If we join these constraints into equation (25), we get

$$\mathbf{A}\mathbf{u} \leq \mathbf{b}$$

$$\begin{bmatrix} \mathbf{I} \\ -\mathbf{I} \\ \mathbf{T} \\ -\mathbf{T} \\ \mathbf{H}_f \\ -\mathbf{H}_f \end{bmatrix} \mathbf{u} \leq \begin{bmatrix} \Delta\mathbf{u}_{\max} \\ -\Delta\mathbf{u}_{\min} \\ \mathbf{u}_{\max} - \mathbf{u}_{k-1} \\ -\mathbf{u}_{\min} + \mathbf{u}_{k-1} \\ \mathbf{y}_{\max} - \mathbf{y}_{free} \\ -\mathbf{y}_{\min} + \mathbf{y}_{free} \end{bmatrix} \tag{32}$$

where \mathbf{I} is an identity matrix, \mathbf{T} is a lower triangular matrix and \mathbf{H}_f is a square matrix of dimension $[(N \cdot n) \times (N \cdot n)]$. The vector \mathbf{b} on the right side of the equation contains column vector of length $(N \cdot m)$ and $(N \cdot n)$. The vector \mathbf{y}_{free} is the free response of the controlled system

$$\mathbf{y}_{free} = \mathbf{H}_p \Delta\mathbf{u}_p + \mathbf{F}\mathbf{x} \tag{33}$$

RESULTS

This section shows the process control simulation results. The simulations demonstrate functionality and possibilities of the designed state-space predictive controller. All of the simulations were done for the system parameters shown in Table 1. Table 2 shows the controller parameters in case without a time-delay and in case with a time-delay. The presence of the time-delay can pretend a situation when a pipeline is attached between the valve changing the input stream and the tank. The process control simulation was done about one operating point with values: $\bar{h}_1 = 7\text{m}$, $\bar{h}_2 = 5\text{m}$, $\bar{q}_{1f} = 0.4469\text{m}^3/\text{min}$, $\bar{q}_{2f} = 0.2150\text{m}^3/\text{min}$.

The control simulations were also compared by two quadratic criterions for analysis of the control quality. The first criterion, described in equation (34), compares the control increments made in every step and the second criterion, described in equation (35), compares a difference between the reference value and the output value.

$$S_u = \frac{1}{N} \sum_{k=1}^N \Delta u^2(k) \tag{34}$$

$$S_e = \frac{1}{N} \sum_{k=1}^N [w(k) - y(k)]^2 \tag{35}$$

Table 1 : System Parameters

Tank	$D[\text{m}]$	$H[\text{m}]$	$k[\text{m}^3/\text{min}]$
1	4	10	0.316
2	4	10	0.296

Table 2 : Controller Parameters

$d[\text{steps}]$	$T_0[\text{min}]$	$N[\text{steps}]$	λ	δ
0	0.25	20	1	1
10	0.25	20	3	0.1

Figure 2 and Figure 3 show the simulation results when the change of both tank liquid levels is desired.

Constraints setup for Figure 2 and Figure 3:

- Case 1: no constraints.
- Case 2:
 - $u_{1\max} = 1 \text{ m}^3/\text{min}$, $u_{1\min} = 0 \text{ m}^3/\text{min}$
 - $u_{2\max} = 1 \text{ m}^3/\text{min}$, $u_{2\min} = 0 \text{ m}^3/\text{min}$
- Case 3:
 - $u_{1\max} = 1 \text{ m}^3/\text{min}$, $u_{1\min} = 0 \text{ m}^3/\text{min}$
 - $u_{2\max} = 1 \text{ m}^3/\text{min}$, $u_{2\min} = 0 \text{ m}^3/\text{min}$
 - $y_{1\max} = 7.5 \text{ m}$, $y_{1\min} = 6.5 \text{ m}$
 - $y_{2\max} = 5.5 \text{ m}$, $y_{2\min} = 4.5 \text{ m}$

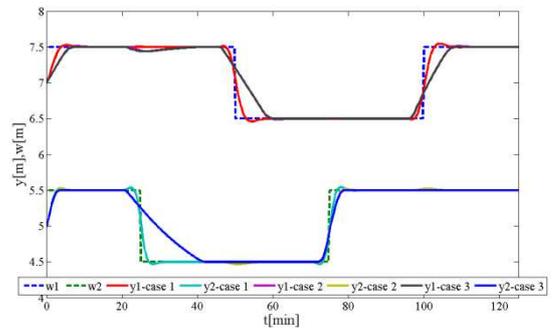


Figure 2 : System output signals

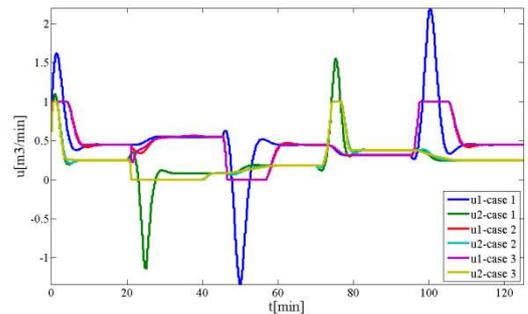


Figure 3 : System control signals

Table 3 : Control simulation results

	Case 1	Case 2	Case 3
$Se_1[m^2]$	0.007	0.021	0.021
$Se_2[m^2]$	0.005	0.023	0.023
$Su_1[(m^3/min)^2]$	$8.61 \cdot 10^{-4}$	$3.85 \cdot 10^{-4}$	$4.03 \cdot 10^{-4}$
$Su_2[(m^3/min)^2]$	$5.24 \cdot 10^{-4}$	$2.83 \cdot 10^{-4}$	$2.85 \cdot 10^{-4}$

Figure 4 and Figure 5 show the simulation results when liquid level of tank 2 is changing and liquid level of tank 1 should stay constant.

Constraints setup for Figure 4 and Figure 5:

- Case 1: no constraints.
- Case 2:
 - $u_{1max} = 1 \text{ m}^3/\text{min}, u_{1min} = 0 \text{ m}^3/\text{min}$
 - $u_{2max} = 1 \text{ m}^3/\text{min}, u_{2min} = 0 \text{ m}^3/\text{min}$
- Case 3:
 - $u_{1max} = 1 \text{ m}^3/\text{min}, u_{1min} = 0 \text{ m}^3/\text{min}$
 - $u_{2max} = 1 \text{ m}^3/\text{min}, u_{2min} = 0 \text{ m}^3/\text{min}$
 - $y_{1max} = 7 \text{ m}, y_{1min} = 7 \text{ m}$

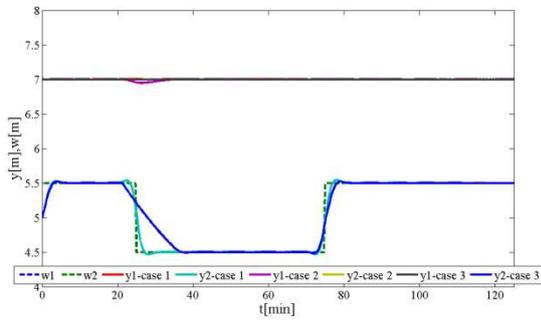


Figure 4 : System output signals

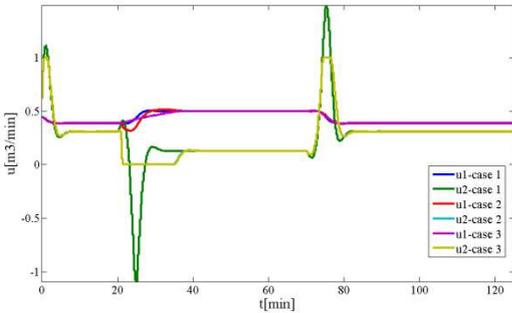


Figure 5 : System control signals

Table 4 :Control simulation results

	Case 1	Case 2	Case 3
$Se_1[m^2]$	$1.91 \cdot 10^{-7}$	$8.89 \cdot 10^{-5}$	$2.26 \cdot 10^{-13}$
$Se_2[m^2]$	0.005	0.016	0.017
$Su_1[(m^3/min)^2]$	$8.26 \cdot 10^{-5}$	$8.37 \cdot 10^{-5}$	$8.30 \cdot 10^{-5}$
$Su_2[(m^3/min)^2]$	$5.21 \cdot 10^{-4}$	$2.89 \cdot 10^{-4}$	$2.89 \cdot 10^{-4}$

Figure 6 and Figure 7 show the simulation results when the 10 steps time-delay is present.

Constraints setup for Figure 6 and Figure 7:

- Case 1: no constraints.
- Case 2:
 - $u_{1max} = 1 \text{ m}^3/\text{min}, u_{1min} = 0 \text{ m}^3/\text{min}$
 - $u_{2max} = 1 \text{ m}^3/\text{min}, u_{2min} = 0 \text{ m}^3/\text{min}$

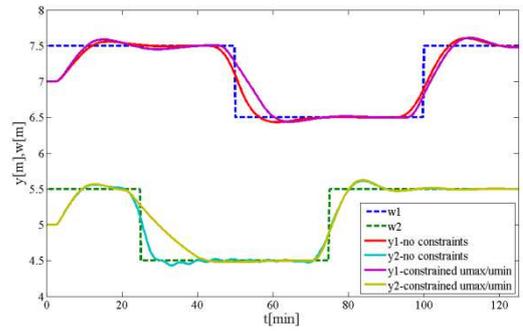


Figure 6 : System output signals with time-delay

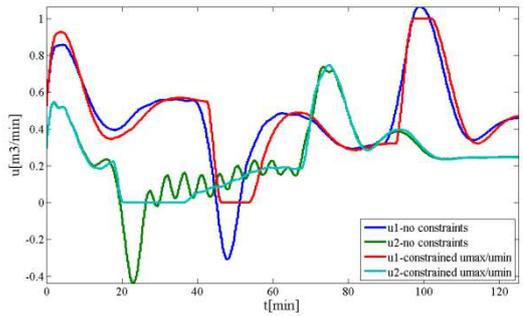


Figure 7 : System control signals with time-delay

Table 5 : Control simulation results

	No constraints	Constrained
$Se_1[m^2]$	0.022	0.031
$Se_2[m^2]$	0.017	0.033
$Su_1[(m^3/min)^2]$	$1.56 \cdot 10^{-4}$	$1.59 \cdot 10^{-4}$
$Su_2[(m^3/min)^2]$	$7.88 \cdot 10^{-5}$	$5.57 \cdot 10^{-5}$

The presented figures show functionality of the constraints of the process variables in the predictive control. The predictive controller without constraints tries to minimize the difference between the reference signal and the output signal and the result can be negative value of the input signal. In this case it means negative liquid stream to the tank. Which is not possible. The possibility of process variables constraints brings a huge advantage into this case. This possibility is directly implemented into the input signal calculation, so there is no need of an additional limitation method. It allows us to keep the input liquid streams between 0 m³/min and 1 m³/min. The constraint of the process variables is not restricted only at one kind of constraint, but they can be combined as Figures 2 to 5 show as case 3. The Figures 4 and 5 show the advantage of the combined constraints when the constant liquid level of the tank 1 is required and the tank 2 liquid level changes. If there is only the absolute control signal value limitation, the liquid level of the first tank drops under the reference value. However, this liquid level stays at the reference value when the hard output signal constraint is combined with the control signal value constraint. The Figures 6 and 7 show another possibility of control process influence. The weighting coefficients

λ and δ are also an option to affect this process. As can be seen, it is necessary to slow down the control process by changing these coefficients to handle it stable.

CONCLUSION

In this paper, the predictive controller based on the state-space CARIMA model with additional possibility of the process variables constraints was presented. This method was used to control the nonlinear process about the selected operating point. The multi-input multi-output system of the two funnel liquid tanks on series was chosen as the exemplar process. The parameters of the tanks simulate large tanks used in an industry. However, the chosen predictive control method works only with linear processes, so the linearization of the nonlinear process is also described in the mathematical model of the controlled system section. One of the main problems of controlling any system is possibility that some process variables can reach a value that is not physically or technologically achievable. The predictive control is a great method to implement the desired process variables constraints directly into the control signal calculation. The results section demonstrates this feature. There are two options how to influence the control process. The first is direct process variables constraints and the second is change of the weighting coefficients λ and δ . This method is also able to control this system with a time-delay with a proper setting of the controller.

REFERENCES

- Albertos Pérez P. and Sala A. 2004. *Multivariable Control Systems: an Engineering Approach*. Springer. London.
- Bars R.; R. Haber and U. Schmitz. 2011. *Predictive control in process engineering: From the basics to the applications*. Weinheim: Wiley-VCH Verlag.
- Bobál, V. 2008, *Adaptive and predictive control*. vol. 1. Zlín, Tomas Bata University in Zlín.
- Camacho E.F. and C. Bordons. 2004. *Model predictive control*, Springer Verlag, London.
- Camacho E.F. and J.E. Normey-Rico. 2007. *Control of dead-time processes*, Springer-Verlag, London.
- Fikar M. and J. Míkleš. 2008. *Process modelling, optimisation and control*, Springer-Verlag, Berlin.
- Hangos K.M.; Bokor J. and Szederkényi G. 2004. *Analysis and Control of Nonlinear Process Systems*. Springer. London.
- Krhovják A.; Dostál P.; Talaš S. and Rušar L. 2015. "Multivariable Gain Scheduled Control of Two Funnel Liquid Tanks in Series". In: International Conference on Process Control 2015, pp. 60-65. Štrbské Pleso. Slovakia
- Maciejowski J.M. 2002. *Predictive control with constraints*, Prentice Hall, London.
- Richardson S.M. 1989. *Fluid Mechanics*. Hemisphere Pub. Corp. New York.
- Rossiter J.A. 2003. *Model based predictive control: a practical approach*, CRC Press.
- Wang L. 2009. *Model predictive control system design and implementation using MATLAB*, Springer Verlag, London.

ACKNOWLEDGMENT

This article was created with support of the Ministry of Education of the Czech Republic under grant IGA reg. n. IGA/FAI/2016/006.

AUTHOR BIOGRAPHIES

LUKÁŠ RUŠAR studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2014. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interests focus on model predictive control. His e-mail address is rusar@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, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His e-mail address is talas@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 interests focus on modeling and simulation of continuous time technological processes, adaptive and nonlinear control.

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, time-delay systems and CAD for automatic control systems. You can contact him on email address bobal@fai.utb.cz.

CASCADE CONTROL OF A TUBULAR CHEMICAL REACTOR USING NONLINEAR PART OF PRIMARY CONTROLLER

Petr Dostal, Jiri Vojtesek, Vladimir Bobal
Department of Process Control, Faculty of Applied Informatics
Tomas Bata University in Zlin
Nad Stranemi 4511, 760 05 Zlin, Czech Republic
E-mail: {dostal;vojtesek;bobal}@fai.utb.cz

KEYWORDS

Cascade control, tubular chemical reactor, external linear control, pole assignment.

ABSTRACT

The paper deals with the cascade control of a tubular chemical reactor. The control is performed in the primary and secondary control-loops. A gain of the primary discrete nonlinear P controller consists of two parts. The first nonlinear part is determined on the basis of simulated or measured steady-state characteristics of the reactor, the second part is selectable. The controller in the secondary control-loop is an adaptive controller. The proposed method is verified by control simulations on a nonlinear model of the reactor with an exothermic reaction.

INTRODUCTION

The cascade control belongs to useful control methods for many technological processes. It may be applied in such cases when at least two output variables can be measured and only one input variable is available to the control. Principles and examples of the use of cascade control can be found e.g. in (Smuts 2011; King 2010; Seborg et al. 1989).

Chemical reactors are typical processes suitable for the use of cascade control. In cases of non-isothermal reactions, concentrations of the reaction products mostly depend on a temperature of the reactant. Further, it is known that while the reactant temperature can be measured almost continuously, concentrations are usually measured in longer time intervals. Then, the application of the cascade control method can lead to good results. In this paper, the procedure for control design of a tubular chemical reactor is presented.

Tubular chemical reactors (TCRs) are units frequently used in chemical industry inclusive of manufacturing and processing of polymers and some others. From the system theory point of view, TCRs belong to the class of nonlinear distributed parameter systems. Their mathematical models are described by sets of nonlinear partial differential equations (PDEs). The methods of modelling and simulation of such processes are described e.g. in (Corriou 2004; Ingham et al. 1994). Detailed analysis of the specific TCR is carried out for example in (Dostál et al. 2008).

In this paper, the TCR control strategy is based on the fact that a concentration of the main component of the reaction taking place in the reactor depends on the output reactant temperature. Moreover, the procedure assumes that the output reactant temperature is measured continuously. Then, in the cascade control-loop, the concentration of a main product of the reaction is considered as the primary controlled variable, and, the output reactant temperature as the secondary controlled variable. The coolant flow rate represents a common control input. The primary controller determining the set point for the secondary (inner) control-loop is derived as a proportional controller with a nonlinear part obtained from the steady-state characteristics of the reactor and with a selectable part. Since the controlled process is nonlinear, a continuous-time adaptive controller is used as the secondary controller. The procedure for the adaptive control design in the inner control-loop is based on approximation of a nonlinear model of the TCR by a continuous-time external linear model (CT ELM) with recursively estimated parameters. In the process of the parameter estimation, a corresponding delta model is used, see, e.g. (Garnier and Wang 2008; Mukhopadhyay et al. 1992; Stericker and Sinha 1993; Bobál et al. 2005). The resulting CT controller is derived on the basis of the polynomial method, see, e.g. (Kučera 1993; Mikleš and Fikar 2004; Dostál et al. 2007). The control is tested by simulations on nonlinear model of the TCR with a consecutive exothermic reaction.

MODEL OF TCR

An ideal plug-flow tubular chemical reactor with a simple exothermic consecutive reaction $A \xrightarrow{k_1} B \xrightarrow{k_2} C$ in the liquid phase and with the countercurrent cooling is considered. 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 reactor model is described by five PDEs in the form

$$\frac{\partial c_A}{\partial t} + v_r \frac{\partial c_A}{\partial z} = -k_1 c_A \quad (1)$$

$$\frac{\partial c_B}{\partial t} + v_r \frac{\partial c_B}{\partial z} = k_1 c_A - k_2 c_B \quad (2)$$

$$\frac{\partial T_r}{\partial t} + v_r \frac{\partial T_r}{\partial z} = \frac{Q_r}{(\rho c_p)_r} - \frac{4U_1}{d_1(\rho c_p)_r} (T_r - T_w) \quad (3)$$

$$\frac{\partial T_w}{\partial t} = \frac{4}{(d_2^2 - d_1^2)(\rho c_p)_w} [d_1 U_1 (T_r - T_w) + d_2 U_2 (T_c - T_w)] \quad (4)$$

$$\frac{\partial T_c}{\partial t} - v_c \frac{\partial T_c}{\partial z} = \frac{4n_1 d_2 U_2}{(d_3^2 - n_1 d_2^2)(\rho c_p)_c} (T_w - T_c) \quad (5)$$

with initial conditions

$$c_A(z, 0) = c_A^s(z), \quad c_B(z, 0) = c_B^s(z), \quad T_r(z, 0) = T_r^s(z),$$

$$T_w(z, 0) = T_w^s(z), \quad T_c(z, 0) = T_c^s(z)$$

and boundary conditions

$$c_A(0, t) = c_{A0}(t) \text{ (kmol/m}^3\text{)}, \quad c_B(0, t) = c_{B0}(t) \text{ (kmol/m}^3\text{)},$$

$$T_r(0, t) = T_{r0}(t) \text{ (K)}, \quad T_c(L, t) = T_{cL}(t) \text{ (K)}.$$

Here, t is the time, z is the axial space variable, c stands for concentrations, T for temperatures, v for fluid velocities, d for diameters, ρ for densities, c_p for specific heat capacities, U for heat transfer coefficients, n_1 is the number of tubes and L is the length of tubes. The subscript $(\cdot)_r$ stands for the reactant mixture, $(\cdot)_w$ for the metal walls of tubes, $(\cdot)_c$ for the coolant, and the superscript $(\cdot)^s$ for steady-state values.

The reaction rates and heat of reactions are nonlinear functions expressed as

$$k_j = k_{j0} \exp\left(\frac{-E_j}{RT_r}\right), \quad j = 1, 2 \quad (6)$$

$$Q_r = (-\Delta H_{r1}) k_1 c_A + (-\Delta H_{r2}) k_2 c_B \quad (7)$$

where k_0 are pre-exponential factors, E are activation energies, $(-\Delta H_r)$ are reaction enthalpies in the negative consideration and R is the gas constant.

The fluid velocities are calculated via the reactant and coolant flow rates as

$$v_r = \frac{4q_r}{\pi n_1 d_1^2}, \quad v_c = \frac{4q_c}{\pi (d_3^2 - n_1 d_2^2)} \quad (8)$$

The TCR parameter values with correspondent units used for simulations are given in Table 1.

From the system engineering point of view, $c_A(L, t) = c_{Aout}$, $c_B(L, t) = c_{Bout}$, $T_r(L, t) = T_{rout}$ and $T_c(0, t) = T_{cout}$ are the output variables and $q_r(t)$, $q_c(t)$, $c_{A0}(t)$, $T_{r0}(t)$ and $T_{cL}(t)$ are the input variables. Among them, for the control purposes, mostly the coolant flow rate can be taken into account as the control variable, whereas other inputs entering into the process can be accepted as disturbances. In this paper, the output reactant temperature

$$T_{rout} = T_r(L, t) \quad (9)$$

is considered as the secondary (inner) controlled output. The concentration c_{Bout} represents the primary controlled output.

TABLE I.: VARIABLES USED IN CONTROL AND APPROXIMATIONS

$L = 8 \text{ m}$	$n_1 = 1200$
$d_1 = 0.02 \text{ m}$	$d_2 = 0.024 \text{ m}$
$d_3 = 1 \text{ m}$	
$\rho_r = 985 \text{ kg/m}^3$	$c_{pr} = 4.05 \text{ kJ/kg K}$
$\rho_w = 7800 \text{ kg/m}^3$	$c_{pw} = 0.71 \text{ kJ/kg K}$
$\rho_c = 998 \text{ kg/m}^3$	$c_{pc} = 4.18 \text{ kJ/kg K}$
$U_1 = 2.8 \text{ kJ/m}^2\text{s K}$	$U_2 = 2.56 \text{ kJ/m}^2\text{s K}$
$k_{10} = 5.61 \cdot 10^{16} \text{ 1/s}$	$k_{20} = 1.128 \cdot 10^{18} \text{ 1/s}$
$E_1/R = 13477 \text{ K}$	$E_2/R = 15290 \text{ K}$
$(-\Delta H_{r1}) = 5.8 \cdot 10^4 \text{ kJ/kmol}$	$(-\Delta H_{r2}) = 1.8 \cdot 10^4 \text{ kJ/kmol}$
$c_{A0}^s = 2.85 \text{ kmol/m}^3$	$c_{B0}^s = 0 \text{ kmol/m}^3$
$T_{r0}^s = 323 \text{ K}$	$T_{cL}^s = 293 \text{ K}$
$q_r^s = 0.15 \text{ m}^3/\text{s}$	

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. The procedure is in detail described in (Dostál et al. 2008).

THE CONTROL OBJECTIVE AND STEADY-STATE CHARACTERISTICS

Basic scheme of the cascade control is in Fig. 1. Here, NPC stands for the nonlinear proportional controller, AC for the adaptive controller.

The control objective is to achieve a concentration of the component B as the primary controlled output near to its maximum. A dependence of the concentration of B on the output reactant temperature is in Fig. 2.

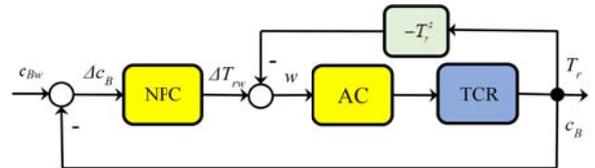


Figure 1: Cascade Control Scheme.

There, an operating interval consists of two parts. In the first subinterval, the concentration B increases with increasing reactant temperature, in the second subinterval it again decreases. The endpoints defining both intervals are

$$315.55 \leq T_{rout} \leq 328.49 \quad 1.344 \leq c_B^s \leq 2.2$$

in the first interval, and,

$$331.32 \leq T_{rout} \leq 334.63 \quad 1.356 \leq c_B^s \leq 2.2$$

in the second interval.

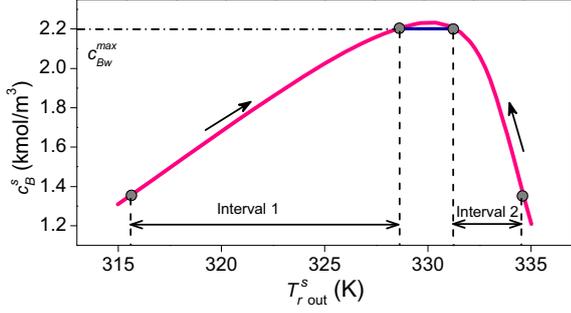


Figure 2: Steady-State Dependence of Product B Concentration on Output Reactant Temperature.

It can be seen in Fig. 2 that the maximum value of c_B can be slightly higher than 2.2 kmol/m^3 . However, the maximum desired value of c_B will be limited just by 2.2 kmol/m^3 .

For purposes of later approximations, the output temperature is transformed as

$$\xi = \frac{T_{r,out} - T_{r,out}^{\min}}{T_{r,out}^{\max} - T_{r,out}^{\min}} \quad (10)$$

where $T_{r,out}^{\min} = 315.46$ and $T_{r,out}^{\max} = 335.01$.

The dependence of c_B on ζ is shown in Fig. 3.

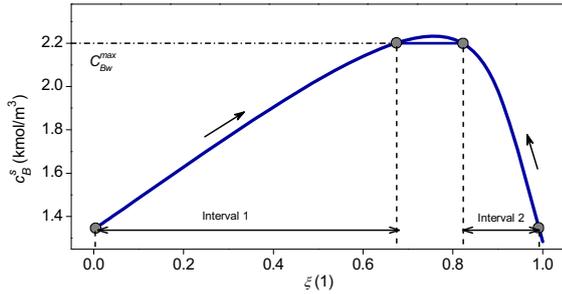


Figure 3: Steady-State Dependence of c_B on ζ .

THE NPC DESIGN

The procedure of the NPC design appears from steady-state characteristics and its subsequent polynomial approximation. Steady-state characteristics with polynomial approximations corresponding with Fig. 3 can be seen in Figs. 4 and 5.

The polynomials for both interval have forms

$$c_B = 1.342 + 1.393\xi + 0.337\xi^2 - 0.747\xi^3 \quad (11)$$

in the first interval, and

$$c_B = -12.247 + 36.386\xi - 22.874\xi^2 \quad (12)$$

in the second interval.

Evidently, the desired value of the reactant output temperature in the output of the NPC can be computed for each c_B as

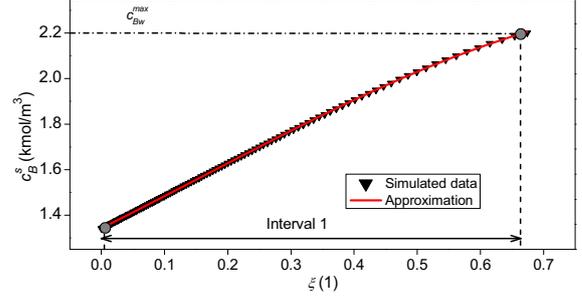


Figure 4: Steady-State Characteristics in the First Interval.

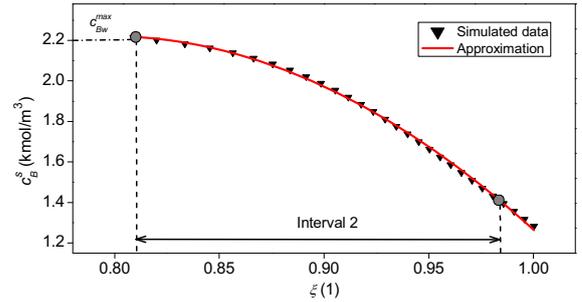


Figure 5: Steady-state Characteristics in the Second Interval.

$$\Delta T_{r,w} = G_w (T_{r,out}^{\max} - T_{r,out}^{\min}) \left(\frac{dc_B}{d\xi} \right)_{c_B}^{-1} \Delta c_{B,w} \quad (13)$$

where $\Delta c_{B,w} = c_{B,w} - c_B$ and G_w is a selectable gain coefficient.

The derivatives of approximate polynomials calculated from (11) and (12) take forms

$$\frac{dc_B}{d\xi} = 1.393 + 0.674\xi - 2.241\xi^2 \quad (14)$$

$$\frac{dc_B}{d\xi} = 36.386 - 45.748\xi \quad (15)$$

The derivatives of approximate polynomials can be seen in Figs. 6 and 7.

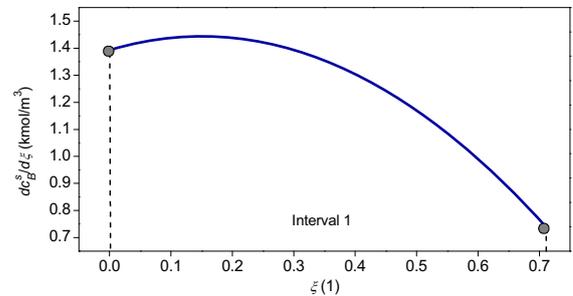


Figure 6: Approximate Polynomial Derivative in the First Interval.

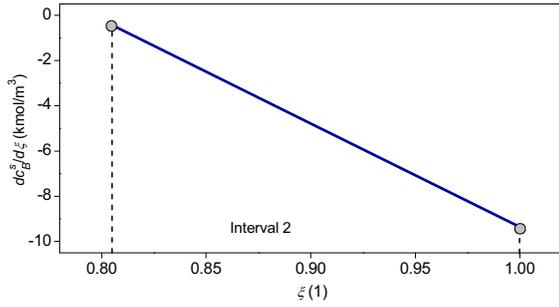


Figure 7: Approximate Polynomial Derivative in the Second Interval.

ADAPTIVE CONTROL SYSTEM DESIGN

Nonlinearity of the reactor is evident from the shape of the steady-state dependence of the output reactant temperature on the coolant flow rate shown in Fig. 8.

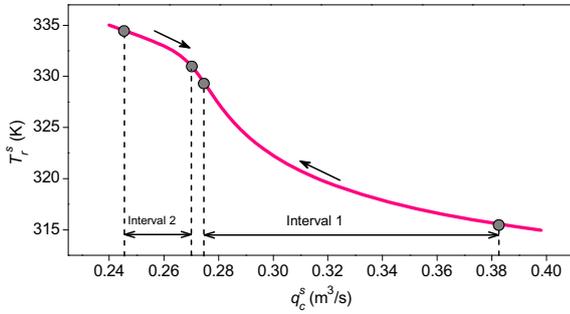


Figure 8: Dependence of the Output Reactant Temperature on the Coolant Flow Rate.

Note that all of the following control simulations are performed in intervals shown in Fig.8.

External Linear Model of the TCR

For the control purposes, the controlled output and the control input are defined as

$$u(t) = \Delta q_c(t) = q_c(t) - q_c^s \quad (16)$$

$$y(t) = \Delta T_{rout}(t) = T_{rout}(t) - T_{rout}^s \quad (17)$$

The CT ELM is proposed in the time domain on the basis of preliminary simulated step responses in the form of the second order differential equation

$$\ddot{y}(t) + a_1 \dot{y}(t) + a_0 y(t) = b_0 u(t) \quad (18)$$

and, in the complex domain, as the transfer function

$$G(s) = \frac{b_0}{s^2 + a_1 s + a_0} \quad (19)$$

External Delta Model

Establishing the δ operator

$$\delta = \frac{d-1}{T_0} \quad (20)$$

where δ is the forward shift operator and T_0 is the sampling period, the delta ELM corresponding to (18) takes the form

$$\delta^2 y(t') + a_1' \delta y(t') + a_0' y(t') = b_0' u(t') \quad (21)$$

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 (21) can be rewritten to the form

$$\delta^2 y(k-2) + a_1' \delta y(k-2) + a_0' y(k-2) = b_0' u(k-2). \quad (22)$$

Then, establishing the regression vector

$$\Phi_\delta^T(k-1) = (-\delta y(k-2) \quad -y(k-2) \quad u(k-2)) \quad (23)$$

where

$$\delta y(k-2) = \frac{y(k-1) - y(k-2)}{T_0} \quad (24)$$

the vector of delta model parameters

$$\Theta_\delta^T(k) = (a_1' \quad a_0' \quad b_0') \quad (25)$$

is recursively estimated by the least squares method with exponential and directional forgetting from the ARX model (see, e.g. (Bobál et al. 2005) in the form

$$\delta^2 y(k-2) = \Theta_\delta^T(k) \Phi_\delta(k-1) + \varepsilon(k) \quad (26)$$

where

$$\delta^2 y(k-2) = \frac{y(k) - 2y(k-1) + y(k-2)}{T_0^2}. \quad (27)$$

Adaptive Controller

The feedback control loop is depicted in Fig. 9. In the scheme, w is the reference signal, e denotes the tracking error, u the control input, and y the controlled output. The transfer function $G(s)$ of the CT ELM is given by (19). The reference w is considered as a step function.

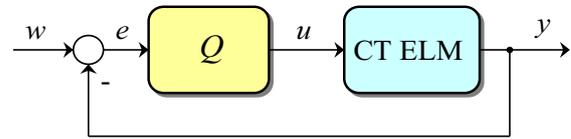


Figure 9: Feedback Control loop.

The feedback controller design is based on the polynomial approach. Procedure for designing can be briefly described as follows:

The transfer function of the controller is in the form

$$Q(s) = \frac{q(s)}{p(s)} \quad (28)$$

where q and p are coprime polynomials satisfying the condition of properness $\deg q(s) \leq \deg p(s)$.

As known, the problem is solved by controller whose polynomials are given by a solution of the polynomial equation

$$a(s)p(s)+b(s)q(s)=d(s) \quad (29)$$

with a stable polynomial $d(s)$ on the right side, with roots representing poles of the closed-loop, and where $p(s)=s\bar{p}(s)$ for step references.

In this paper, the polynomial $d(s)$ is considered in the form

$$d(s)=n(s)(s+\alpha)^2 \quad (30)$$

where the polynomial $n(s)$ is chosen as a result of spectral factorization

$$a^*(s)a(s)=n^*(s)n(s) \quad (31)$$

and α is a selectable double pole.

For $G(s)$ with $a(s)=s^2+a_1s+a_0$, the resulting controller has the transfer function

$$Q(s)=\frac{q_2s^2+q_1s+q_0}{s(s+p_0)} \quad (32)$$

with parameters computed from (26).

The above procedure implies that the controller parameters can be adjusted by the single selectable parameter α .

Simulation Experiments

Simulations document an effect of the concentration c_B measurement period and the adjustable part of the NPC gain on the control signals.

All simulations were performed on nonlinear model of the TCR. In the secondary control-loop, the P controller with a small gain was used at the start of simulations. For the δ -model parameter recursive identification, the sampling period $T_0=0.5$ s was chosen. The value of the selectable parameter α is stated under each figure.

In the first case, simulations in the first operating interval started from the point $c_B^s=1.3504$ kmol/m³, $T_{rout}^s=315.55$ K and $q_c^s=0.383$ m³/s. The desired value of c_B has been chosen as $c_{Bw}=2.2$ kmol/m³. An effect of the parameter G_w on the reference w , output temperature T_{rout} , the concentration c_B and the control input q_c responses is evident from Figs. 10 – 13.

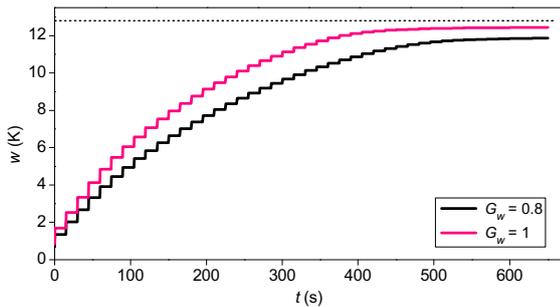


Figure 10: Reference Signal Courses ($t_s=15$, $\alpha=0.05$).

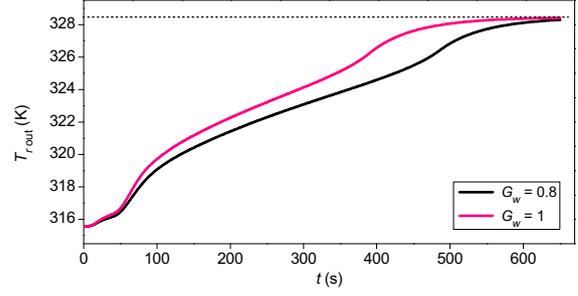


Figure 11: Reactant output temperature responses ($t_s=15$, $\alpha=0.05$).

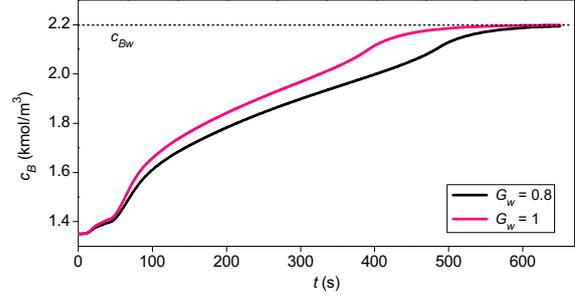


Figure 12: Concentration c_B responses ($t_s=15$, $\alpha=0.05$).

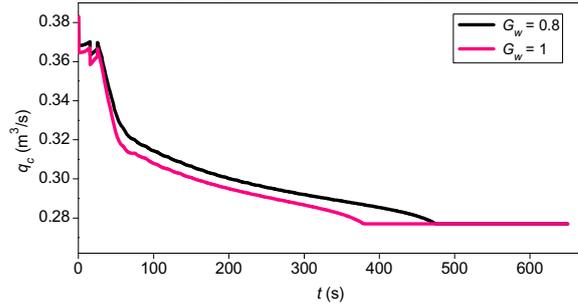


Figure 13: Coolant flow rate responses ($t_s=15$, $\alpha=0.05$).

It can be seen that an increasing G_w accelerates all signals in the control loop. However, its value is not unrestricted and its convenient value should be found experimentally.

An effect of the period t_s in the same operating interval can be seen in Figs.14 – 17. Although shortening t_s leads to faster control responses, its length is determined by possibilities of measurement.

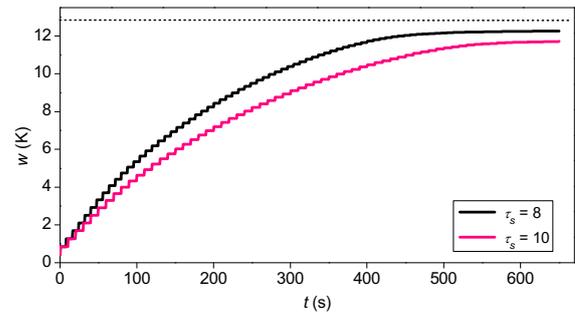


Figure 14: Reference Signal Courses ($G_w=0.5$, $\alpha=0.05$).

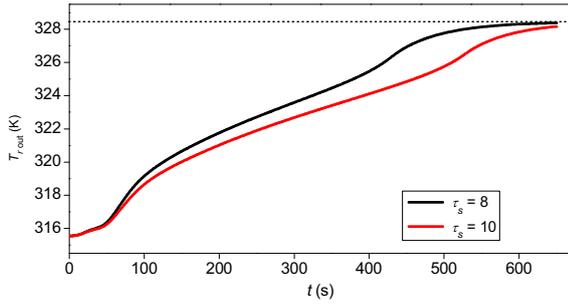


Figure 15: Reactant Output Temperature Responses ($G_w = 0.5$, $\alpha = 0.05$).

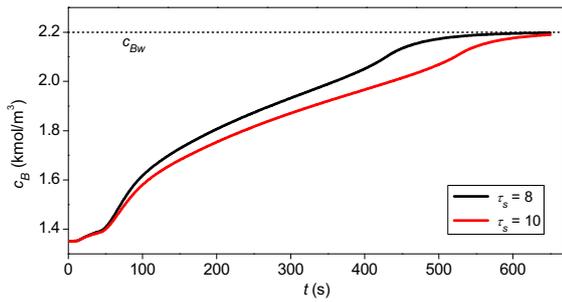


Figure 16: Concentration c_B responses ($G_w = 0.5$, $\alpha = 0.05$).

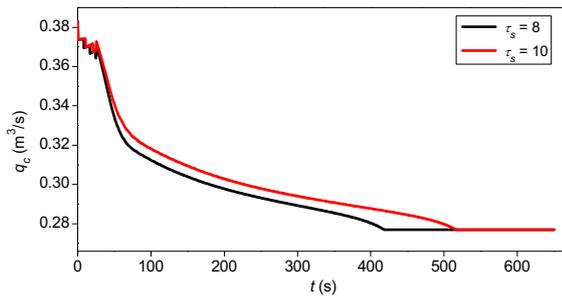


Figure 17: Coolant flow rate responses ($G_w = 0.$, $\alpha = 0.05$).

In the second case, simulations in the second operating interval started from the point $c_B^s = 1.3564$ kmol/m³, $T_{r,out}^s = 334.63$ K and $q_c^s = 0.244$ m³/s. The desired value of c_B has again been chosen as $c_{Bw} = 2.2$ kmol/m³. Here, only signal courses for $G_w = 1.2$, $t_s = 10$ and $\alpha = 0.08$ are presented. All signal courses are shown in Figs. 18 – 21.

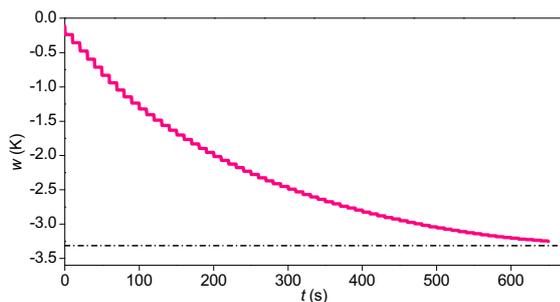


Figure 18: Reference Signal Course.

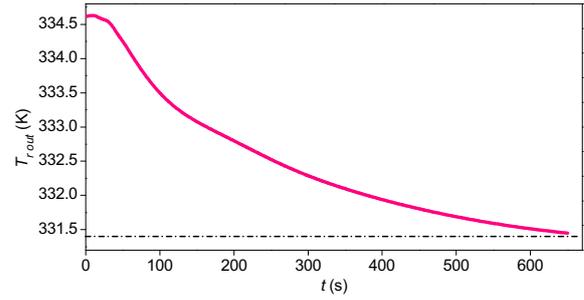


Figure 19: Reactant Output Temperature Response.

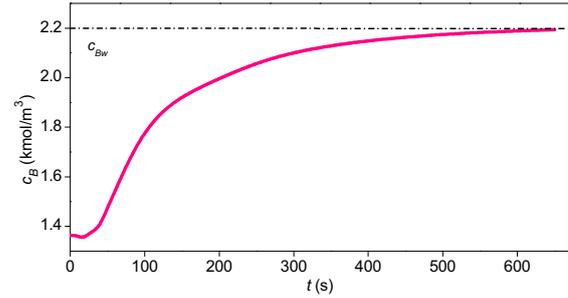


Figure 20: Concentration c_B response.

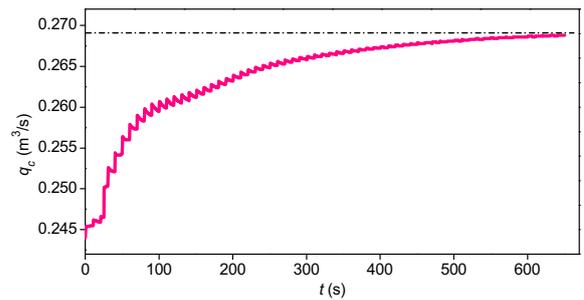


Figure 21: Coolant flow rate response.

CONCLUSIONS

The paper deals with design of the cascade control of a tubular chemical reactor. The presented procedure supposes measuring both the output concentration of a main reaction product and the output reactant temperature. The control is performed in the external and inner closed-loop where the concentration of a main product is the primary and an output reactant temperature the secondary controlled variable. A common control input is the coolant flow rate.

The controller in the external control-loop is a discrete nonlinear P-controller derived on the basis of simulated or measured steady-state characteristics of the reactor. The controller in the inner control-loop is an adaptive continuous-time controller. In its derivation, the recursive parameter estimation of an external delta model of the reactor, the polynomial approach and the pole placement method are applied.

The presented method has been tested by computer simulation on the nonlinear model of the tubular chemical reactor with a consecutive exothermic reaction.

REFERENCES

- Bobál, V., J. Böhm, J. Fessl, and J. Macháček. 2005. *Digital self-tuning controllers*, Springer Verlag, Berlin.
- Corriou, J.-P. 2004. *Process control. Theory and applications*. Springer – Verlag, London.
- Dostál, P., F. Gazdoš, V. Bobál, and J. Vojtěšek. 2007. "Adaptive control of a continuous stirred tank reactor by two feedback controllers". In *Proc. 9th IFAC Workshop Adaptation and Learning in Control and Signal Processing ALCOSP'2007*, Saint Petersburg, Russia, P5-1 – P5-6.
- Dostál, P., V. Bobál, and J. Vojtěšek. 2008. "Simulation of steady-state and dynamic behaviour of a tubular chemical reactor". In *Proc. 22nd European Conference on Modelling and Simulation*, Nicosia, Cyprus, 487-492.
- Ingham, J., I.J. Dunn, E. Heinzle, and J.E. Přenosil (1994). *Chemical Engineering Dynamics. Modelling with PC Simulation*, VCH Verlagsgesellschaft, Weinheim.
- Garnier, H. and L. Wang (eds.). 2008. *Identification of continuous-time models from sampled data*. Springer-Verlag, London.
- King, M. (2010). *Process Control: A Practical Approach*, John Wiley, Chichester, UK.
- Kučera, V. 1993. "Diophantine equations in control – A survey". *Automatica*, 29, 1361-1375.
- Mikleš, J., and M. Fikar. (2004). *Process modelling, identification and control 2*, STU Press, Bratislava.
- 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.
- Seborg, D.E., T.F. Edgar, and D.A. Mellichamp. 1989. *Process dynamics and control*. John Wiley and Sons, Chichester.
- Smuts, J.F. 2011. *Process control for practitioners*. OptiControls, New York.
- 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 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.



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. In the year 2015 he became associate professor. He now works 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.



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.

CONTINUOUS-TIME VS. DISCRETE-TIME IDENTIFICATION MODELS USED FOR ADAPTIVE CONTROL OF NONLINEAR PROCESS

Jiri Vojtesek and Petr Dostal
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

Simulation, Mathematical Model, Adaptive control, Continuous-time model, Delta-model, Continuous Stirred-tank Reactor.

ABSTRACT

An adaptive control is a technique where the controller adopts a structure or parameters somehow to the control conditions and the state of the controlled system. One way how we can fulfil the adaptivity of the controller is a recursive identification of the controlled system which satisfies that parameters of the controller changes according to parameters of the controlled system during the whole control process. The goal of this contribution is to compare identification models that work in continuous and discrete time. The control synthesis uses polynomial approach that satisfies basic control requirements such as a stability, a disturbance attenuation and a reference signal tracking. The control response could be tuned by the choice of the root position in the Pole-placement method. Moreover, this control method could be easily programmable that is big advantage while we use this method in simulation software such as Matlab etc.

INTRODUCTION

The adaptive control (Åström and Wittenmark, 1989) is not new control approach but it is still used because it produces good control results. Advantage of this method can be found in very good theoretical background and variety of modifications (Bobal *et al.*, 2005).

The approach used here is based on the choice of the External Linear Model (ELM) which describes controlled, originally nonlinear, process in the linear way for example by the discrete or the continuous transfer function (TF) (Bobal *et al.*, 2005). Parameters of this ELM are then identified recursively during the control and parameters of the controller are recomputed according to them. Results of control synthesis are the structure and relations for computing controller's parameters that reflect identified parameters of ELM.

The recursive identification of the continuous-time (CT) model (Wahlberg, 1990) is a bit more complicated than identification of the discrete-time (DT) model where the computation uses measured or simulated values of input

and output variables in discrete time intervals. This approach could be inaccurate for bigger values of the sampling period. One solution can be found in the use of so called delta-models (Middleton and Goodwin, 2004) that are special types of DT models where parameters of input and output variables are related to the sampling period. It was proved that parameters of the delta-model approach to parameters of the CT model for sufficiently small sampling period (Stericker and Sinha, 1993). This combination of the continuous-time control synthesis with the discrete-time identification is called "Hybrid adaptive control" and some applications can be found for example in (Vojtesek and Dostal, 2005) and (Vojtesek and Dostal, 2011).

The second way is to use the CT control synthesis and also the CT recursive identification. The CT online estimation is not as simple as a DT estimation because derivatives of the input and output variable are immeasurable. This negative feature could be solved for example with the use of differential filters (Dostal *et al.*, 2001).

The control synthesis uses polynomial approach which satisfies basic control system requirements such as a stability of the control loop, a reference signal tracking and a disturbance attenuation. Moreover, the two degrees-of-freedom (2DOF) configuration has good results in the reference signal tracking (Kucera, 1993).

The continuous stirred-tank reactor (CSTR) is typical nonlinear equipment used in the chemical and biochemical industry for production of various chemicals (Ingham *et al.*, 2000). The mathematical model of this nonlinear system is described by the set of nonlinear ordinary differential equations (ODEs) which can be solved mathematically for example by the Runge-Kutta's method. This mathematical model than serves as a testing model for simulation analyses proposed in the theoretical part.

All results in this paper are simulations made in the mathematical software Matlab, version 7.0.1.

ADAPTIVE CONTROL

The adaptive approach (Åström and Wittenmark, 1989) takes its philosophy in the nature, where plants, animals or even human beings "adapt" their behavior to the actual conditions and environment they live in. There could be various adaptive control techniques but the one

which is used in this work adapt parameters of the controller to actual state of the controlled system. This done via recursive identification of the system's ELM and parameters of the controller are then recomputed according to identified parameters of the ELM.

The design of the controller starts with the choice of the ELM. We can use for example transfer functions (TF) that are generally described in the CT form:

$$G(s) = \frac{b(s)}{a(s)} \quad (1)$$

where polynomials $a(s)$ and $b(s)$ will be later used in the computation of controller's parameters.

It is good to do the static and dynamic analysis of the controlled system before the design of the controller. The static analysis helps with the choice of the optimal working point where we can obtain for example the best concentration of the product or minimal costs. On the other hand, the dynamic analysis of the system can be used for example for the choice of the ELM's order.

Continuous-Time Identification Model

As $G(s)$ is also relation of the Laplace transform of the output variable, $Y(s)$, to the input variable, $U(s)$, the ELM in the (1) could be also rewritten to the form

$$a(\sigma) \cdot y(t) = b(\sigma) \cdot u(t) \quad (2)$$

where $u(t)$ denotes the input variable, $y(t)$ is the output variable and σ is the differentiation operator.

The identification of CT model in (2) is problem because the derivatives of the input and the output variables are immeasurable. If we replace these derivations by the filtered ones denoted by u_f and y_f and computed from

$$\begin{aligned} c(\sigma) \cdot u_f(t) &= u(t) \\ c(\sigma) \cdot y_f(t) &= y(t) \end{aligned} \quad (3)$$

for a new stable polynomial $c(\sigma)$ that fulfils condition $\deg c(\sigma) \geq \deg a(\sigma)$, the Laplace transform of (3) is then

$$\begin{aligned} c(s) \cdot U_f(s) &= U(s) + o_1(s) \\ c(s) \cdot Y_f(s) &= Y(s) + o_2(s) \end{aligned} \quad (4)$$

where polynomials $o_1(s)$ and $o_2(s)$ includes initial conditions of filtered variables. If we substitute (4) into the Laplace transform of the Equation (2), the relation for the Laplace transform of the filtered output variable, $Y_f(s)$ is

$$Y_f(s) = \frac{b(s)}{a(s)} U_f(s) + \Psi(s) \quad (5)$$

and $\Psi(s)$ is a rational function which contains initial conditions of both filtered and unfiltered variables.

The dynamics of the differential filters $c(s)$ in (4) must be faster than the dynamics of the controlled system (Dostal *et al.*, 2001). It is good to choose the parameters of this polynomial sufficiently small.

The values of filtered values are taken in the discrete time moment $t_k = k \cdot T_v$ for $k = 0, 1, 2, \dots, N$. T_v is sampling period and the regression vector has $n+m$ parts where $\deg a = n$ and $\deg b = m$, i.e.

$$\begin{aligned} \boldsymbol{\varphi}_{CT}(t_k) &= [-y_f(t_k), -y_f^{(1)}(t_k), \dots, -y_f^{(n-1)}(t_k), \dots \\ &\dots, u_f(t_k), u_f^{(1)}(t_k), \dots, u_f^{(m)}(t_k), 1]^T \end{aligned} \quad (6)$$

The vector of parameters

$$\boldsymbol{\theta}_{CT}(t_k) = [a_0, a_1, \dots, a_{n-1}, b_0, b_1, \dots, b_m]^T \quad (7)$$

is computed from the differential equation

$$y_f^{(n)}(t_k) = \boldsymbol{\theta}_{CT}^T(t_k) \cdot \boldsymbol{\varphi}_{CT}(t_k) + \Psi(t_k) \quad (8)$$

where $\Psi(t_k)$ includes immeasurable errors.

Discrete-Time Identification Model

The second approach used for example in (Vojtesek and Dostal, 2011) uses so called delta-models for identification. The delta-models are special types of DT models where input and output variables are related to the sampling period.

A new complex variable γ is defined generally as (Mukhopadhyay *et al.*, 1992)

$$\gamma = \frac{z-1}{\beta \cdot T_v \cdot z + (1-\beta) \cdot T_v} \quad (9)$$

where T_v denotes a sampling period and β is an optional parameter and it holds $0 \leq \beta \leq 1$. It is clear, that there could be an infinite number of delta-models but so called *Forward delta-model* for $\beta = 0$ was used here.

The complex variable γ is then

$$\gamma = \frac{z-1}{T_v} \quad (10)$$

Some works compares parameters CT vs. delta-model and it was proved for example in (Stericker and Sinha, 1993), that parameters of the delta-model approaches to the CT ones for sufficiently small sampling period T_v .

The CT model (2) can be rewritten to

$$a'(\delta) y(t') = b'(\delta) u(t') \quad (11)$$

where $a'(\delta)$ and $b'(\delta)$ are discrete polynomials and their coefficients are different from those in CT model but we suppose, that they are close to them.

The regression vector is in this case

$$\begin{aligned} \boldsymbol{\varphi}_\delta(k-1) &= [-y_\delta(k-1), \dots, -y_\delta(k-n), \\ &u_\delta(k+m-n), \dots, u_\delta(k-n)]^T \end{aligned} \quad (12)$$

The vector of parameters is generally

$$\boldsymbol{\theta}_\delta(k) = [a'_{n-1}, \dots, a'_0, b'_m, \dots, b'_0]^T \quad (13)$$

and its parameters are computed again from the differential equation

$$y_\delta(k) = \boldsymbol{\theta}_\delta^T(k) \cdot \boldsymbol{\varphi}_\delta(k-1) + e(k) \quad (14)$$

for $e(k)$ as a general random immeasurable component. Both identification methods with the CT model and the delta-model was discussed in this work.

Recursive Identification

Vectors of CT and delta parameters must be identified recursively to satisfy the adaptivity condition. This could be done for example by the Recursive Least-Squares (RLS) method (Fikar and Mikles 1999) which is simple, easily programmable method that could be modified with exponential, directional etc. forgetting factors. These forgetting factors helps with the accuracy in the more complex systems. The RLS method used for estimation of vectors of parameters $\hat{\boldsymbol{\theta}}_{CT}^T$ or $\hat{\boldsymbol{\theta}}_\delta^T$ in (7) and (14) could be described generally by the set of equations:

$$\begin{aligned} \varepsilon(k) &= y(k) - \boldsymbol{\varphi}^T(k) \cdot \hat{\boldsymbol{\theta}}(k-1) \\ \gamma(k) &= [1 + \boldsymbol{\varphi}^T(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k)]^{-1} \\ \mathbf{L}(k) &= \gamma(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k) \\ \mathbf{P}(k) &= \frac{1}{\lambda_1(k-1)} \left[\mathbf{P}(k-1) - \frac{\mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k) \cdot \boldsymbol{\varphi}^T(k) \cdot \mathbf{P}(k-1)}{\lambda_1(k-1) + \boldsymbol{\varphi}^T(k) \cdot \mathbf{P}(k-1) \cdot \boldsymbol{\varphi}(k)} \right] \\ \hat{\boldsymbol{\theta}}(k) &= \hat{\boldsymbol{\theta}}(k-1) + \mathbf{L}(k) \varepsilon(k) \end{aligned} \quad (15)$$

where $\boldsymbol{\varphi}$ is regression vector, ε denotes a prediction error, \mathbf{P} is a covariance matrix and λ_1 and λ_2 are forgetting factors. For example constant exponential forgetting (Fikar and Mikles 1999) uses $\lambda_2 = 1$ and

$$\lambda_1(k) = 1 - K \cdot \gamma(k) \cdot \varepsilon^2(k) \quad (16)$$

where K is a very small value (e.g. $K = 0.001$). This RLS modification was used in this work for the online estimation.

DESIGN OF THE CONTROLLER

The controller is designed with the use of the polynomial synthesis (Kucera, 1993). There are several advantages of this approach. At first, they can work with the controller in the polynomial description, for example in the form of the transfer function (1). The result of the synthesis is not only the structure, but also the relations for computing of controller's parameters. Moreover, this method satisfies basic control requirements.

The control scheme with two degrees-of-freedom (2DOF) (Grimble, 1994) is shown in Figure 1.

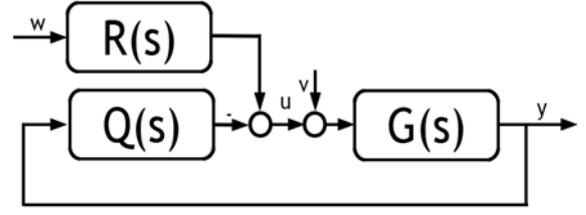


Figure 1: 2DOF control configuration

The signal w is reference signal (i.e. wanted value), v denotes disturbance, u is an input and y an output variable. The block $G(s)$ in Figure 1 represents the controlled system described by the TF (1), blocks $Q(s)$ and $R(s)$ are feedback and feedforward parts of the controller again in the form of TF, generally:

$$Q(s) = \frac{q(s)}{p(s)}; R(s) = \frac{r(s)}{p(s)} \quad (17)$$

Degrees of polynomials $p(s)$, $q(s)$ and $r(s)$ must hold properness condition:

$$\deg q(s) \leq \deg p(s); \deg r(s) \leq \deg p(s) \quad (18)$$

The condition for the reference signal tracking is satisfied if the polynomial $p(s)$ in the denominator of the controller's transfer functions (17) is divided into

$$p(s) = f(s) \cdot \tilde{p}(s) \quad (19)$$

where $f(s)$ is a least common divisor of the reference and the disturbance transfer functions. If we have these TF in the form of the step function, $f(s) = s$ and (17) could be rewritten into

$$Q(s) = \frac{q(s)}{s \cdot \tilde{p}(s)}; R(s) = \frac{r(s)}{s \cdot \tilde{p}(s)} \quad (20)$$

Parameters of controller's polynomials are computed from the set of polynomial equations

$$\begin{aligned} a(s) \cdot s \cdot \tilde{p}(s) + b(s) \cdot q(s) &= d(s) \\ t(s) \cdot s + b(s) \cdot r(s) &= d(s) \end{aligned} \quad (21)$$

that are in the literature called *Diophantine equations* (Kucera, 1993) and they can be solved by the Method of uncertain coefficients. Polynomial $t(s)$ in equation (21) is an auxiliary stable polynomial and coefficients of this polynomial are not used for computing of coefficients of the polynomial $r(s)$.

Polynomials $a(s)$ and $b(s)$ in (21) are known from the recursive identification and the polynomial $d(s)$ on the right side of Diophantine equations (21) is stable optional polynomial which could affect the quality of the control.

Degrees of controller's polynomials $\tilde{p}(s)$, $q(s)$ and $r(s)$ and the degree of the stable polynomial $d(s)$ are

$$\begin{aligned} \deg \tilde{p}(s) &= \deg a(s) - 1 & \deg r(s) &= 0 \\ \deg q(s) &= \deg a(s) & \deg d(s) &= 2 \cdot \deg a(s) \end{aligned} \quad (22)$$

The simplest way how to choose the stable optional polynomial $d(s)$ define the *Pole-placement method* The polynomial $d(s)$ is then divided into

$$d(s) = \prod_{i=1}^{\deg d(s)} (s + s_i) \quad (23)$$

where roots s_i are generally in the complex form $s_i = \alpha_i + \omega_i \cdot j$ and the stability is satisfied for $\alpha_i < 0$. If we want to obtain an aperiodic output response, ω_i must be $\omega_i = 0$ and (23) is then

$$d(s) = (s + \alpha)^{\deg d} \quad (24)$$

One disadvantage of this method is that it is very general and it provides for example for $\deg d(s) = 4$ four simple roots, two double roots, one single and one triple root but no recommendation for the choice of these roots.

Our previous experiment (Vojtesek and Dostal, 2011) have shown, that it is good to connect the choice of this polynomial somehow with the controlled system. The *Spectral factorization* could be used for this task and it means that the polynomial $d(s)$ is divided into two parts

$$d(s) = n(s) \cdot (s + \alpha)^{\deg d - \deg n} \quad (25)$$

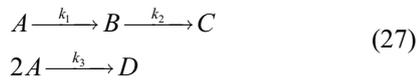
where one part is classic pole-placement method and $n(s)$ comes from the Spectral factorization of the polynomial $a(s)$ in the denominator of the controlled system's transfer function (1):

$$n^*(s) \cdot n(s) = a^*(s) \cdot a(s) \quad (26)$$

The use of Spectral factorization satisfies that the polynomial $n(s)$ is always stable even if the polynomial $a(s)$ is unstable. This could happen for example by inaccurate estimation at the beginning of the control when an estimator does not have enough information about the system.

SIMULATION EXPERIMENT

The adaptive approach was tested by simulations on the mathematical model of the Continuous Stirred-Tank Reactor (CSTR) with so called Van der Vusse reaction inside (Chen *et al.*, 1995). This reaction can be described by the following scheme:



and the mathematical model of this system comes from material and heat balances inside the reactor. The result is the set of four nonlinear ordinary differential equations (ODE):

$$\begin{aligned} \frac{dc_A}{dt} &= \frac{q_r}{V_r} (c_{A0} - c_A) - k_1 c_A - k_3 c_A^2 \\ \frac{dc_B}{dt} &= -\frac{q_r}{V_r} c_B + k_1 c_A - k_2 c_B \end{aligned} \quad (28)$$

$$\begin{aligned} \frac{dT_r}{dt} &= \frac{q_r}{V_r} (T_{r0} - T_r) - \frac{h_r}{\rho_r c_{pr}} + \frac{A_r U}{V_r \rho_r c_{pr}} (T_c - T_r) \\ \frac{dT_c}{dt} &= \frac{1}{m_c c_{pc}} (Q_c + A_r U (T_r - T_c)) \end{aligned} \quad (28)$$

State variables are in this case concentrations c_A , c_B and temperatures of the reactant T_r and the cooling T_c . There could be theoretically four input variables – a volumetric flow rate of the reactant, q_r , a heat removal of the cooling, Q_c , an input concentration c_{A0} and an input temperature of the reactant, T_{r0} . The last two are only theoretical and could not be used as an input variable from the practical point of view.

The scheme of this chemical reactor is in Figure 2.

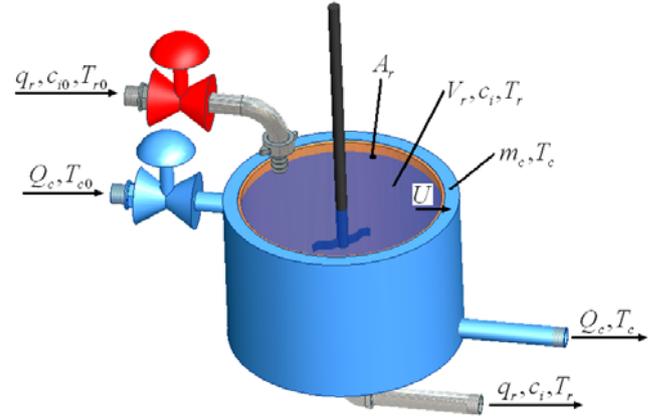


Figure 2: Continuous Stirred-tank Reactor (CSTR) with Van der Vusse reaction inside

Other variables are supposed to be constant during the control because of the simplification. The volume of the reactor is denoted as V_r , A_r is the heat exchange surface, ρ_r is used for the density of the reactant, U is the heat transfer coefficient, c_{pc} and c_{pr} are specific heat capacities of the cooling and the reactant a m_c is the weight of the cooling mass. Values of these fixed parameters are in Table 1 (Chen *et al.*, 1995).

Table 1: Parameters of the reactor

$k_{01} = 2.145 \cdot 10^{10} \text{ min}^{-1}$	$k_{02} = 2.145 \cdot 10^{10} \text{ min}^{-1}$
$k_{03} = 1.5072 \cdot 10^8 \text{ min}^{-1} \text{ mol}^{-1}$	$E_1/R = 9758.3 \text{ K}$
$E_2/R = 9758.3 \text{ K}$	$E_3/R = 8560 \text{ K}$
$h_1 = -4200 \text{ kJ.kmol}^{-1}$	$h_2 = 11000 \text{ kJ.kmol}^{-1}$
$h_3 = 41850 \text{ kJ.kmol}^{-1}$	
$V_r = 0.01 \text{ m}^3$	$\rho_r = 934.2 \text{ kg.m}^{-3}$
$c_{pr} = 3.01 \text{ kJ.kg}^{-1} \text{.K}^{-1}$	$c_{pc} = 2.0 \text{ kJ.kg}^{-1} \text{.K}^{-1}$
$U = 67.2 \text{ kJ.min}^{-1} \text{m}^{-2} \text{K}^{-1}$	$A_r = 0.215 \text{ m}^2$
$c_{A0} = 5.1 \text{ kmol.m}^{-3}$	$c_{B0} = 0 \text{ kmol.m}^{-3}$
$T_{r0} = 387.05 \text{ K}$	$m_c = 5 \text{ kg}$

The steady-state analysis (Vojtesek and Dostal, 2005) has shown that the optimal working point is in this case defined by the volumetric flow rate of the reactant $q_r^s = 2.4 \cdot 10^{-3} \text{ m}^3 \text{.min}^{-1}$ and heat removal of the coolant $Q_c^s = -18.56 \text{ kJ.min}^{-1}$.

The input variable for the dynamic study was the change of the heat removal of the coolant, ΔQ_c , and the output variable was the change of reactant's temperature, T_r ,

$$u(t) = \frac{Q_c(t) - Q_c^s}{Q_c^s} \cdot 100 [\%], \quad y(t) = T(t) - T^s [K] \quad (29)$$

The dynamic behavior was observed for various step changes of the input variable from the range $\Delta u(t) = \langle -100\%; +100\% \rangle$ and results are in Figure 3.

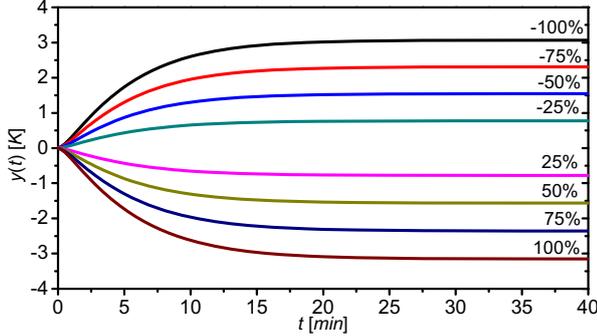


Figure 3: Results of the dynamic analysis for the various changes of the input variable $\Delta u(t)$

It was already mentioned, that the dynamic analysis could help us with the choice of the ELM. It can be seen, that resulted step responses could be approximated by the second order TF with relative order one. The TF (1) is then

$$G(s) = \frac{b(s)}{a(s)} = \frac{b_1 s + b_0}{s^2 + a_1 s + a_0} \quad (30)$$

As the ELM (30) is of the second order, the TF of the controller for both identification methods are according to (20) and (22)

$$Q(s) = \frac{q_2 s^2 + q_1 s + q_0}{s \cdot (p_1 s + p_0)}; \quad R(s) = \frac{r_0}{s \cdot (p_1 s + p_0)} \quad (31)$$

and the stable polynomial $d(s)$ on the right side of (21) is of the fourth degree, i.e.

$$d(s) = n(s) \cdot (s + \alpha)^2 \quad (32)$$

where $n(s)$ comes from the Spectral factorization of (26) Finally, we have one tuning parameter – the position of the root α .

All simulations have same parameters – the sampling period was $T_v = 0.3 \text{ min}$, the initial covariance matrix $\mathbf{P}(0)$ has on the diagonal $1 \cdot 10^6$ and starting vectors of parameters for the identification was chosen $\hat{\theta}_{CT}(0) = \hat{\theta}_s(0) = [0.1, 0.1, 0.1, 0.1]^T$. The simulation took 750 min and there were done 5 changes of the reference signal $w(t)$. Our previous experiments have shown that we can obtain better control results if the first change of the reference signal is exponential function instead of the step function. The input signal

$u(t)$ was limited to the values $u(t) = \langle -75\%; +75\% \rangle$ due to physical limitations.

Control with CT Identification Model

The first simulation experiment was done for CT identification model. The degree of the polynomial $c(s)$ was chosen as $\deg c(s) = \deg a(s) = 2$ and

$$c(s) = s^2 + c_1 s + c_0 = s^2 + 1.4s + 0.49 \quad (33)$$

Filtered input and output variables are then

$$\begin{aligned} y_f^{(2)}(t) + c_1 y_f^{(1)}(t) + c_0 y_f(t) &= y(t) \\ u_f^{(2)}(t) + c_1 u_f^{(1)}(t) + c_0 u_f(t) &= u(t) \end{aligned} \quad (34)$$

The vector of parameters and the regression vector are for ELM (30)

$$\begin{aligned} \varphi_{CT}(t_k) &= [-y_f(t_k), -y_f^{(1)}(t_k), u_f(t_k), u_f^{(1)}(t_k)]^T \\ \theta_{CT}(t_k) &= [a_0, a_1, b_0, b_1]^T \end{aligned} \quad (35)$$

where parameters $\theta_{CT}(t_k)$ are estimated recursively by RLS method with constant exponential forgetting described in the theoretical part.

There were done three simulation studies for different α and results are shown in Figure 4 and Figure 5.

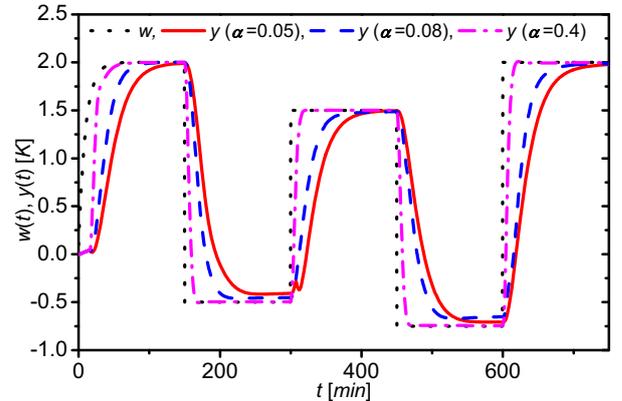


Figure 4: Courses of the reference signal, $w(t)$, and the output variable, $y(t)$, for various values of the parameter α , results for the CT identification model

Figure 4 clearly shows that increasing value of the parameter α affects mainly the speed of the output response – an increasing value of α produces quicker output response. It is worth to notice, that the change of the reference signal from the positive to the negative value causes problems for smaller values of α . The output response then do not reach the reference signal. On the other hand, the control with the biggest value of $\alpha = 0.4$ produces very good results also with this negative step changes of the reference signal.

Figure 5 shows that the controller computes also very smooth course of the action value, $u(t)$, what is also important from the practical point of view. The action signal is represented by some action of the actuators and

quick changes of the input variable could affect the lifetime of them.

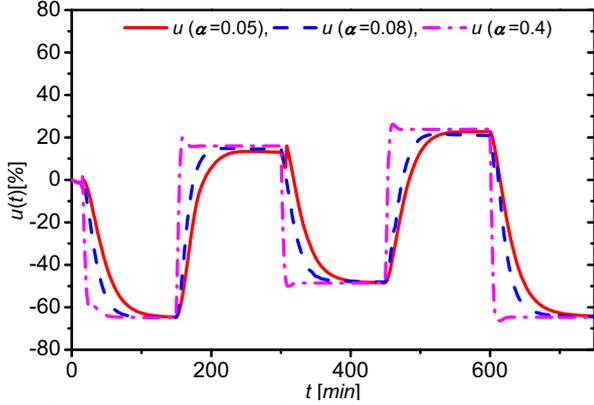


Figure 5: The course of the input variable, $u(t)$, for various values of the parameter α , results for the CT identification model

Control with Delta-model Identification

The second approach uses delta-models for identification which means that vector of parameters and data vector are

$$\begin{aligned} \boldsymbol{\varphi}_\delta(k-1) &= [-y_\delta(k-1), -y_\delta(k-2), u_\delta(k-1), u_\delta(k-2)]^T \\ \boldsymbol{\theta}_\delta(k) &= [a'_1, a'_0, b'_1, b'_0]^T \end{aligned} \quad (36)$$

where δ -values of the input and the output variables are

$$\begin{aligned} y_\delta(k) &= \frac{y(k) - 2y(k-1) + y(k-2)}{T_v^2} \\ y_\delta(k-1) &= \frac{y(k-1) - y(k-2)}{T_v} \\ y_\delta(k-2) &= y(k-2) \\ u_\delta(k-1) &= \frac{u(k-1) - u(k-2)}{T_v} \\ u_\delta(k-2) &= u(k-2) \end{aligned} \quad (37)$$

Parameters of $\hat{\boldsymbol{\theta}}_\delta(k)$ are again estimated recursively by the RLS method with constant exponential forgetting.

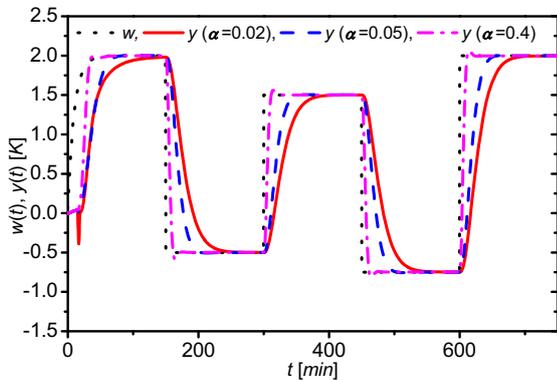


Figure 6: Courses of the reference signal, $w(t)$, and the output variable, $y(t)$, for various values of the parameter α , results for the delta identification model

There were done simulation experiments for the same values of the parameter $\alpha = 0.05, 0.08$ and 0.4 and the same changes of the reference signal as in previous case due to comparability. Results are shown in Figure 6 and Figure 7.

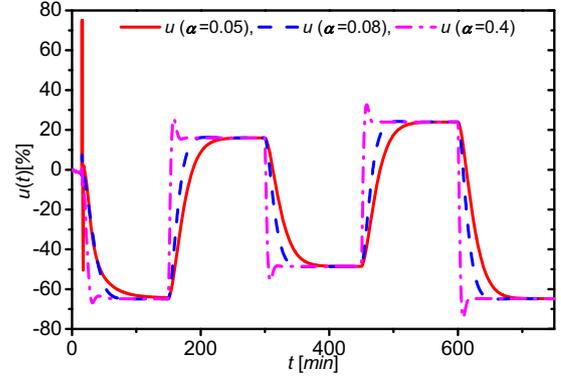


Figure 7: The course of the input variable, $u(t)$, for various values of the parameter α , results for the delta identification model

It can be seen that the use of delta-models for the identification can also produce good control results. The effect of α is the same as in previous case, i.e. quicker output response can be obtained for bigger values of α . The output response in this case have a very small overshoots for the biggest value of $\alpha = 0.4$ but does not have problem with negative changes of the reference signal compared with the CT model. The course of the input variable, $u(t)$, is very similar to the previous case. We can see only some problems at the very beginning of the control which is typical for the adaptive control that starts from the general vector of parameters $\hat{\boldsymbol{\theta}}_\delta(0)$. It takes some time to approach to right values, but once they are reached, results are good. The quality of the control for both control techniques was evaluated by the control quality criteria S_u that displays how big are changes of the input variable $u(t)$ and the output criteria S_y that sums the square of the control error $e = w - y$, i.e.

$$\begin{aligned} S_u &= \sum_{i=2}^N (u(i) - u(i-1))^2 [-] \\ S_y &= \sum_{i=1}^N (w(i) - y(i))^2 [K^2] \end{aligned} \quad \text{for } N = \frac{T_f}{T_v} \quad (38)$$

where T_f is final time which is in this case $T_f = 450 \text{ min}$. Values of these quality criteria was computed for each simulation study and results are shown in Table 2.

Table 2: Values of quality criteria S_u and S_y

	CT model		Delta model	
	$S_u [-]$	$S_y [K^2]$	$S_u [-]$	$S_y [K^2]$
$\alpha = 0.02$	278	1 254	23 939	1 733
$\alpha = 0.05$	374	886	390	1 233
$\alpha = 0.4$	2 203	312	1 704	453

The choice of the optimal value of the tuning parameter α in both strategies depends what is important for us from the control point of view. Table 2 shows that if the output variable is more important, the bigger value of α is better. This can be also clearly seen from graphs in Figure 4 and Figure 6. Oppositely, if we want the less changes of the input variable, the control with lower value of α is good choice.

As all results in this paper comes from the simulation it is worth to mention the computation requirements. The simulation of the control with the delta identification model takes in Matlab about 10 *seconds*. On the other hand, the CT identification model is more computationally demanding and the simulation for the same parameters took 1.5 *minutes*. As a result, computation with CT model is nine times more demanding than control with delta identification model. In fact it is not big problem because the sampling period was $T_s = 0.3 \text{ min}$, which is 20 *seconds* that is enough time for the identification and the computation of new parameters of the controller even for CT model.

CONCLUSIONS

The goal of this paper was to show two on-line recursive identification methods used in the adaptive control. The first one is continuous-time identification that uses differential filters. This method is more computationally demanding but offers more accurate results. The next identification method is based on delta-models that are special types of DT models where input and output variables are related to the sampling period which could shift parameters of the delta-model close to parameters of the CT model. As a result, this method is quicker and the output responses are very close to those from the CT model. Used adaptive approach uses polynomial approach with the Pole-placement method and the Spectral factorization that satisfies basic control requirements. Moreover, this adaptive controller could be tuned by the choice of the position of the root in the Pole-placement method and the main effect is in the speed of the control. All approaches were tested on the mathematical model of the CSTR as a typical member of the nonlinear systems with lumped parameters. The future work will head to the verification of simulated results on the real model of this or similar system.

REFERENCES

- Åström, K. J., B. Wittenmark 1989. *Adaptive Control*. Addison Wesley. Reading, MA.
- Bobal, V.; J. Böhm; J. Fessl; 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.
- Dostal, P.; V. Bobal; M. Blaha, M. 2001. One Approach to Adaptive Control of Nonlinear Processes. In: *Proc. IFAC Workshop on Adaptation and Learning in Control and*

- Signal Processing ALCOSP 2001*, Cernobbio-Como, Italy, 2001. p. 407-412.
- Fikar, M.; J. Mikles 1999. *System Identification*. STU Bratislava
- Grimble, M. J. 1994. *Robust industrial control. Optimal design approach for polynomial systems*. Prentice Hall, London. 1994.
- Chen, H.; A. Kremling; F. Allgöwer. 1995. Nonlinear Predictive Control of a Benchmark CSTR. In: *Proceedings of 3rd European Control Conference*. Rome, Italy.
- 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.
- Kucera, V. 1993. Diophantine Equations in Control – A Survey. *Automatica*. 29. 1361-1375
- Middleton, R.H.; G. C. Goodwin 2004. *Digital Control and Estimation - A Unified Approach*. Prentice Hall. Englewood Cliffs.
- Mukhopadhyay, S.; A. G. Patra; G. P. Rao. 1992. New Class of Discrete-time Models for Continuous-time Systems. *International Journal of Control*, 1992, vol.55, 1161-1187.
- Stericker, D.L; N. K. Sinha 1993. Identification of Continuous-time Systems from Samples of Input-output Data Using the δ -operator. *Control-Theory and Advanced Technology*. vol. 9. 113-125
- Vojtesek, J.; P. Dostal, 2005. From steady-state and dynamic analysis to adaptive control of the CSTR reactor. In: *Proc. of 19th European Conference on Modelling and Simulation ECMS 2005*. Riga, Latvia, p. 591-598
- Vojtesek, J.; P. Dostal. 2011. Two Types of External Linear Models Used for Adaptive Control of Continuous Stirred Tank Reactor. In: *Proceedings 25th European Conference on Modelling and Simulation ECMS 2011*. Nicosia: p. 501-507. ISBN 978-0-9564944-2-9.
- Wahlberg, B. 1990. The Effects of Rapid Sampling in System Identification, *Automatica*, vol. 26, 167-170.

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 and become Associative professor at the Tomas Bata University in Zlin in 2015. His email contact is vojtesek@fai.utb.cz.



PETR DOSTAL studied at the Technical University of Pardubice. He obtained his PhD. degree in Technical Cybernetics in 1979 and he became professor in Process Control in 2000. His research interest are modelling 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.

OPTIMAL GAIN SCHEDULED CONTROLLER FOR A TWO FUNNEL LIQUID TANKS IN SERIES

Adam Krhovják, Petr Dostál, Stanislav Talaš and Lukáš Rušar
Department of Process Control
Faculty of Applied Informatics, Tomas Bata University in Zlín,
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
krhovjak@fai.utb.cz

KEYWORDS

Two funnel liquid tanks, nonlinear model, parametrized linear model, scheduling variable, integral control, optimal gain scheduled controller.

ABSTRACT

Motivated by the control complexity of nonlinear systems, we introduce an optimal gain scheduled controller for a nonlinear system of two funnel liquid tanks in series, based on a linearization of a nonlinear state equation of the system about selected operating points. Specifically, the proposed technique aims at extending the region of validity of linearization by introducing a parametrized linear model, which enables to construct a feedback controller at each point. Additionally, we present an integral control approach which ensures robust regulation under all parameter perturbations. The parameters of resulting family of feedback controllers are scheduled as functions of reference variables, resulting in a single optimal controller. Nonlocal performance of the gain scheduled controller for the nonlinear model has been checked by mathematical simulation.

INTRODUCTION

We live in a world of highly complex systems that exhibit nonlinear behavior. Engineers facing the control of these systems are required to design such mechanisms that would satisfy desired characteristics through the operating range. In many cases, situation is complicated by the fact that the tracking problem involves multiple variables interacting with each other. One consequence is that superposition principle, which is known from linear systems does not hold any longer and we are faced more challenging situations.

However, because of powerful tools we know for linear systems, the first step in designing a control for a nonlinear systems consists in linearization. There is no question that whenever it is possible, we should take advantage of design via linearization approach.

Nevertheless, we must bear in mind the basic limitation associated with this approach. The understanding that linearization represents only an approximation in the neighborhood of an operating point.

To put it differently, linearization cannot be viewed globally since it can only predict the local behavior of a nonlinear system.

Therefore, conventional controllers with fixed parameters cannot guarantee performance beyond the vicinity of operating point. Interestingly enough, in many situation, it is possible to explicitly capture how the dynamic of the system changes in its equilibrium points by introducing a family of linear models which are parametrize by one or more scheduling variables. In such cases it is intuitively reasonable to linearize the nonlinear system about selected operating points, capturing key states of the system, design a linear feedback controller at each point, and interpolate the resulting family of linear controllers by monitoring scheduling variables. There has been a significant research in gain scheduling (GS). We refer the interest reader to (Rugh 1991; Shamma and Athans 1991; Shamma and Athans 1992; Lawrence and Rugh 1995) for deeper and more insightful understanding of the gain scheduling procedure.

Most efforts have been devoted to the analytical framework (Shamma and Athans 1990; Rugh 1991) and less attention has been paid to particular engineering applications except few remarkable applications in technological processes (Jiang 1994; Kaminer et al. 1995; Krhovják et al. 2015). Moreover, these intensive efforts have not paid attention to the question of optimal performance design which would make the concept much more attractive for a potential practitioner in industry. In order to address those needs we have stressed to illustrate gain scheduling strategy for a nonlinear system of two funnel liquid tanks in series (TFLT), extending region of validity of linearization approach by designing an optimal controller that is a prescription for moving from one design to another.

Thus, the problem of a designing an optimal control for a model of the nonlinear system of the TFLT has been reduced to a problem of designing a family of optimal feedback controllers that are interpreted as a single controller via scheduling variables.

Throughout the paper we gradually reveal the scheduling procedure satisfying a tracking problem as well as the design of an optimal control trajectory for the multivariable nonlinear system of two funnel tanks.

MODEL OF THE TFLT

A simplified model of the TFLT system taken from (Dostál et al. 2008) is schematically shown in Figure 1. The process consists of two liquid streams that are pumped into funnel tanks. Pump with a flow rate q_{1f} discharges liquid into the first tank (T1). The second tank

(T2) is fed by both liquid stream q_{2f} and q_1 , representing liquid that leaves the first tank through the opening in the base. There are no reactants or reaction kinetics and stoichiometry to consider. The model also includes hydraulic relationship for the tank outlet streams.

Both parameters of the tanks and initial liquid levels are captured in Table 1.

Table 1: Model parameters

Tank	D	H	\bar{h}
	m	m	m
1	1.5	2.5	1.8
2	1.5	2.5	14

In this case study we have used $k_1 = 0.32 \text{ m}^{2.5}/\text{min}$, $k_2 = 0.3 \text{ m}^{2.5}/\text{min}$ and $q_{1f} = 0.2 \text{ m}^3/\text{min}$, $q_{2f} = 0.1 \text{ m}^3/\text{min}$.

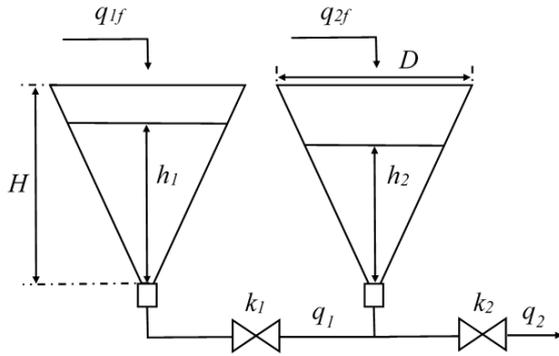


Figure 1: Two funnel liquid tanks in series

The only step needed to develop the model of SLT is to write conservation equation (Richardson 1989), representing material balance for a single material. Recall that the general form of a mass balance is given

INPUT = OUTPUT + ACCUMULATION

It is easy to see that the simplified model of TFLT can be modeled by

$$\pi \frac{D^2}{4H^2} h_1^2 \frac{dh_1}{dt} + q_1 = q_{1f} \quad (1)$$

$$\pi \frac{D^2}{4H^2} h_2^2 \frac{dh_2}{dt} - q_1 + q_2 = q_{2f} \quad (2)$$

where h_1 and h_2 represent liquid levels from the bottom and D is the diameter of the cross sectional area at the top of the tanks. As the liquid moves through the valves, we see dependence of q on liquid level as

$$q_1 = k_1 \sqrt{|h_1 - h_2|} \quad (3)$$

$$q_2 = k_2 \sqrt{h_2} \quad (4)$$

where k_1 and k_2 are positive valve constants.

MODEL STRUCTURE FOR GS DESIGN

In the process of designing and implementing a gain scheduled controller for a nonlinear system, we have to find its approximations about the family of operating

(equilibrium) points. Thus, the coupled nonlinear first-order ordinary differential equations (1)-(2) capturing the dynamics of the TFLT have to be transformed into its linearized form.

In view of our example, we shall deal with multi-input multi-output linearizable nonlinear system represented by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \quad (5)$$

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \quad (6)$$

where $\dot{\mathbf{x}}$ denotes derivative of \mathbf{x} with respect to time variable and \mathbf{u} are specified input variables. We call the variable \mathbf{x} the state variable and \mathbf{y} the output variable. We shall refer to (5) and (6) together as the state-space model.

To obtain a state-space model of the TFLT, let us take $x_1 = h_1$, $x_2 = h_2$ as state variables and $u_1 = q_{1f}$, $u_2 = q_{2f}$ as control inputs. Then the state equations are

$$\frac{dx_1}{dt} = \frac{4H^2}{\pi D^2 x_1^2} (u_1 - k_1 \sqrt{|x_1 - x_2|}) \quad (7)$$

$$\frac{dx_2}{dt} = \frac{4H^2}{\pi D^2 x_2^2} (u_2 + k_1 \sqrt{|x_1 - x_2|} - k_2 \sqrt{x_2}) \quad (8)$$

and the output equations take the form

$$y_1 = x_1 \quad (9)$$

$$y_2 = x_2 \quad (10)$$

One can easily sketch the trajectory of steady-state characteristic by setting $\dot{\mathbf{x}} = \mathbf{0}$ and solving for unknown vector \mathbf{x} .

Therefore the equilibrium points correspond to the solution of

$$0 = \frac{4H^2}{\pi D^2 x_1^2} (u_1 - k_1 \sqrt{|x_1 - x_2|}) \quad (11)$$

$$0 = \frac{4H^2}{\pi D^2 x_2^2} (u_2 + k_1 \sqrt{|x_1 - x_2|} - k_2 \sqrt{x_2}) \quad (12)$$

Having calculated equilibrium points of state equation, our goal now is to approximate (5) about selected single operating point. Suppose $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{u} \neq \mathbf{0}$, and consider the change of variables

$$y_1 = x_1 \quad (13)$$

$$y_2 = x_2 \quad (14)$$

$$y_1 = x_1 \quad (15)$$

It should be noted that in the new variables system has equilibria in origin.

Expanding the right hand side of (5) about point $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$, we obtain

$$\mathbf{f}(\mathbf{x}, \mathbf{u}) \approx \mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}}) + \frac{\partial \mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}})}{\partial \mathbf{x}} (\mathbf{x} - \bar{\mathbf{x}}) + \frac{\partial \mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}})}{\partial \mathbf{u}} (\mathbf{u} - \bar{\mathbf{u}}) + \text{H.O.T.} \quad (16)$$

If we restrict our attention to a sufficiently small neighborhood of the equilibrium point such that the

higher-order terms are negligible, then we may drop these terms and approximate the nonlinear state equation by the linear state equation

$$\dot{\mathbf{x}}_\delta = \mathbf{A}\mathbf{x}_\delta + \mathbf{B}\mathbf{u}_\delta \quad (17)$$

where

$$\mathbf{A} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\bar{\mathbf{x}}, \mathbf{u}=\bar{\mathbf{u}}} \quad (18)$$

$$\mathbf{B} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right|_{\mathbf{x}=\bar{\mathbf{x}}, \mathbf{u}=\bar{\mathbf{u}}}$$

PARAMETRIZATION OF LINEAR MODELS

Before we present a parametrization via scheduling variable, let us first examine configuration of the gain scheduled control system captured in Figure 2. From the figure, it can be easily seen that controller parameters are automatically changed in open loop fashion by monitoring operating conditions. From this point of view, presented gain scheduled control system can be understand as a feedback control system in which the feedback gains are adjusted using feedforward gain scheduler.

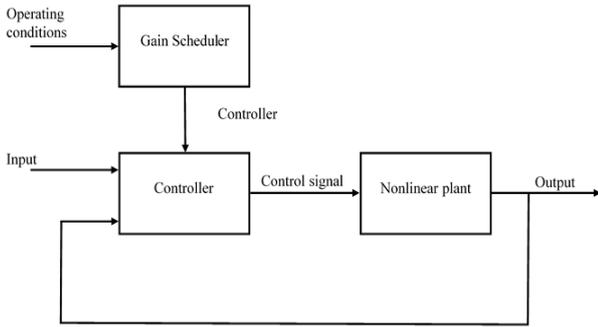


Figure 2: Gain scheduled control

Then it comes as no surprise that first and the most important step in designing a controller is to find an appropriate scheduling strategy. Once the strategy is found, it can be directly embedded into the controller design.

In order to understand the idea behind the gain scheduling let us first consider the nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \boldsymbol{\alpha}) \quad (19)$$

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \boldsymbol{\alpha}) \quad (20)$$

We can see that the nonlinear system is basically same as the system that we have introduced in the previous section by equations (5) and (6). The only difference here is that both state and output equations are parameterized by a new *scheduling variable* $\boldsymbol{\alpha}$ representing the operating conditions.

To illustrate this motivating discussion let us consider this crucial point in the context of our example.

Suppose the system is operating at steady state and we want to design controller such that \mathbf{x} tracks a reference signal \mathbf{w} . In order to maintain the output of the plant at the

value \bar{x}_1 and \bar{x}_2 , we have to generate the corresponding input signal to the system at $\bar{u}_1 = k_1 \sqrt{x_1 - x_2}$ and $\bar{u}_2 = k_2 \sqrt{x_2}$, respectively. This implies that for every value of \mathbf{w} in the operating range, we can define the desired operating point by $\mathbf{x} = \mathbf{y} = \mathbf{w}$ and $\mathbf{u} = \bar{\mathbf{u}}(\mathbf{w})$

Thus it means, that we can directly schedule on a reference trajectory.

Having identified a scheduling variable, the common scheduling scenario takes this form

$$\dot{\mathbf{x}}_\delta = \mathbf{A}(\boldsymbol{\alpha})\mathbf{x}_\delta + \mathbf{B}(\boldsymbol{\alpha})\mathbf{u}_\delta \quad (21)$$

Intuitively speaking, the parameters of (21) are scheduled as functions of the scheduling variable $\boldsymbol{\alpha}$. Since our model is simple nonlinear TITO system, we need to calculate elements of \mathbf{A} , \mathbf{B} , corresponding to the structure of (18). In other words, the key how to move from one operating point to another is given by

$$a_{01}(\boldsymbol{\alpha}) = \frac{4H^2 k_1 \sqrt{\alpha_1 - \alpha_2}}{2D^2 \pi \alpha_1^2 (\alpha_1 - \alpha_2)}$$

$$a_{02}(\boldsymbol{\alpha}) = -\frac{4H^2 k_1 \sqrt{\alpha_1 - \alpha_2}}{2D^2 \pi \alpha_1^2 (\alpha_1 - \alpha_2)}$$

$$a_{03}(\boldsymbol{\alpha}) = -\frac{4H^2 k_1 \sqrt{\alpha_1 - \alpha_2}}{2D^2 \pi \alpha_2^2 (\alpha_1 - \alpha_2)} \quad (22)$$

$$a_{04}(\boldsymbol{\alpha}) = \frac{4H^2}{2D^2 \pi \alpha_2^2} \left[\frac{k_1 \sqrt{\alpha_1 - \alpha_2}}{\alpha_1 - \alpha_2} + \frac{k_2 \sqrt{\alpha_2}}{\alpha_2} \right]$$

$$b_{01}(\boldsymbol{\alpha}) = \frac{4H^2}{D^2 \pi \alpha_1^2}, \quad b_{04}(\boldsymbol{\alpha}) = \frac{4H^2}{D^2 \pi \alpha_2^2}$$

An important feature of our analysis is that even if $\boldsymbol{\alpha}$ represents reference vector, the equations (22) still capture the behavior of the system around equilibria.

REGULATION VIA INTEGRAL CONTROL

Since the previous sections resulted in a family of parametrized linear models we want to design state feedback control such that

$$\mathbf{y} \rightarrow \mathbf{y}_r \text{ as } t \rightarrow \infty \quad (23)$$

Further, we assume that we can physically measure the controlled output \mathbf{y} . In order to ensure zero steady-state tracking error in the presence of uncertainties, we want to use integral control. The regulation task will be achieved by stabilizing system at an equilibrium point where $\mathbf{y} = \mathbf{y}_r$.

To maintain the system at that point it must be true, that there exists a pair of $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ such that

$$\mathbf{0} = \mathbf{f}(\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\boldsymbol{\alpha}}) \quad (24)$$

$$\mathbf{0} = \mathbf{g}(\bar{\mathbf{x}}, \bar{\boldsymbol{\alpha}}) - \mathbf{y}_r \quad (25)$$

Note, that for equations (24)-(25) we assume a unique solution $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$.

Toward the goal, we have integrate the tracking error

$$e = y - y_r \quad (26)$$

$$\dot{\sigma} = e \quad (27)$$

Having defined the integrator of the tracking error let now augment the system (19) to obtain

$$\dot{x} = f(x, u, \alpha) \quad (28)$$

$$\dot{\sigma} = g(x, \alpha) - y_r \quad (29)$$

It follows the control u will be designed as a feedback function of (x, σ) . For such control the new system has an equilibrium point $(\bar{x}, \bar{\sigma}, \bar{\alpha})$.

To proceed with the design of the controller, we now linearize (28)-(29) about $(\bar{x}, \bar{\sigma}, \bar{\alpha})$ to obtain augmented state space model as

$$\dot{\xi} = \begin{bmatrix} A(\alpha) & \mathbf{0} \\ C(\alpha) & \mathbf{0} \end{bmatrix} \xi + \begin{bmatrix} B(\alpha) \\ \mathbf{0} \end{bmatrix} v \stackrel{\text{def}}{=} A(\alpha)\xi + B(\alpha)v \quad (30)$$

where

$$\xi = \begin{bmatrix} x - \bar{x} \\ \sigma - \bar{\sigma} \end{bmatrix}, v = u - \bar{u} \quad (31)$$

Now we have to design a matrix $K(\alpha)$ such that $A - BK$ is Hurwitz.

Partition $K(\alpha)$ as $K(\alpha) = -[K_1(\alpha) \ K_2(\alpha)]$ implies that the state feedback control should be taken as

$$u = -K_1(\alpha)(x - \bar{x}) - K_2(\alpha)(\sigma - \bar{\sigma}) + \bar{u} \quad (32)$$

and by applying the control (32), we obtain the closed-loop system

$$\dot{x} = f(x, -K_1(\alpha)(x - \bar{x}) - K_2(\alpha)(\sigma - \bar{\sigma}) + \bar{u}) \quad (33)$$

$$\dot{\sigma} = g(x, \alpha) - y_r \quad (34)$$

Figure 3 clearly illustrates the block diagram of the control system, where we can clearly see embedded integral control action

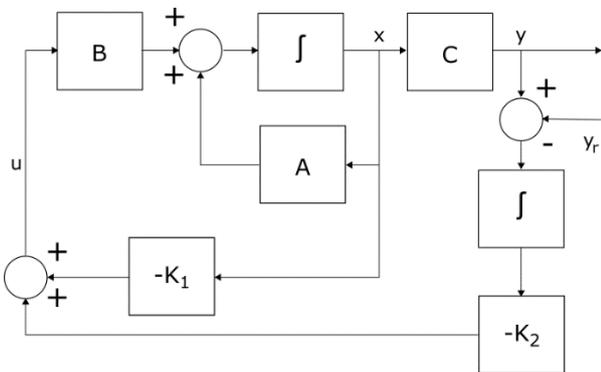


Figure 3: Block diagram of integral control system

In searching for an optimal control $u = -Kx$ we have to design gain matrix K that minimizes quadratic cost function (omit dependence on α)

$$J = \int_0^{\infty} [x^T Q x + u^T R u + 2x^T N u] dt \quad (35)$$

where Q, R a symmetric, positive (semi-) definite weighting matrices and N represents

The traditional problem is solved using algebraic Riccati equation

$$A^T P + P A - P B + N R^{-1} (B^T P + N^T) + Q = \mathbf{0} \quad (36)$$

Then the gain matrix K is derived from P by

$$K = R^{-1} (B^T P + N^T) \quad (37)$$

In view of the procedure that we have just described, one can notice that three main issues are involved in the development of gain scheduled controller; namely linearization of TFLT about the family of operating regions, design of a parametrized family of linear matrix feedback controllers for the parametrized family of linear systems and construction of gain scheduled controller.

So far, we have formed the basic idea of the control problem. All that remains now is to simulate the performance of the gain scheduling procedure with the help of the integral control.

SIMULATIONS AND RESULTS

In this section, we simulate the gain scheduled control of TFLT. We have developed a custom MATLAB function based on the simulator introduced by (Krhovjak et al. 2014) that simulates adequately the behavior of TFLT. Idealistic model has been implemented according to equations (1) and (2). The popular ODE solver using based on Runge-Kutta methods (Hairer et al. 1993) was considered to calculate numerical solution.

The simulation results of gain scheduled control are presented in Figures 4-8. Figure 4 shows the optimal responses of the control system to sequences of step changes in reference signals. As can be seen a step change in reference signals causes a new calculation of the equilibrium point of the system.

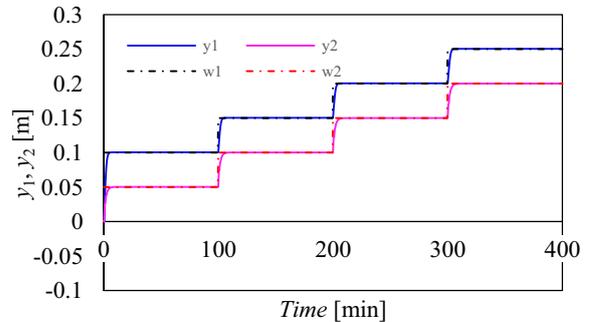


Figure 4: The responses of the closed-loop system to a sequence of step changes

Figure 5 shows the response of the closed-loop system to a slow ramp that takes the set points over a period of 500 minutes. These observations are consistent with a common gain scheduling rule-of-thumb about the behavior of gain scheduled controller under slowly varying scheduling variable.

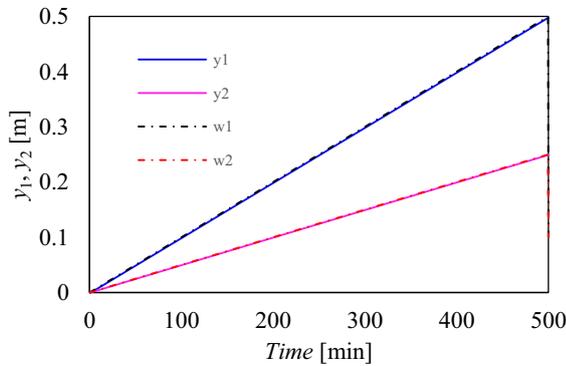


Figure 5: Slow ramp

In contrast, the figure 6 shows the response to a faster ramp signal. As the slope of the ramp increases, tracking performance deteriorates. If we keep increasing the slope of the ramp, the system will eventually go unstable.

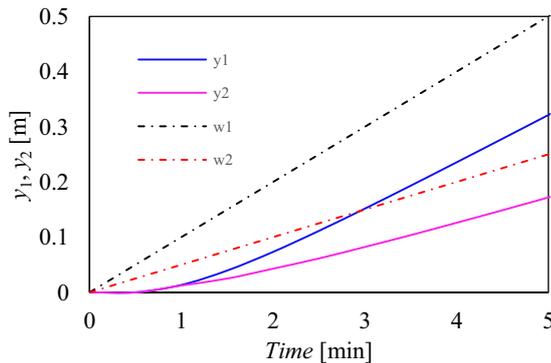


Figure 6: Fast ramp

To appreciate what we gain by gain scheduling, Figure 7 and Figure 8 illustrates responses of the closed-loop system to the same sequence of changes. In the first case, a gain scheduled controller is applied, while in the second case a fixed-gain controller evaluated at $\alpha = [-1 \ -1]$ is used.

From this illustration it is evident why we have to modify the gain scheduled controller. While stability and zero steady-state tracking error are achieved, as predicted by our analysis, the responses deteriorates significantly as the reference is far from operating point. In some situations it may be possible to reach a large value of the reference signal by a sequence of step changes, as in the Figure 4 where we allow enough time for the system to settle done after each step change. This can be viewed as another possible way how to change the reference set point.

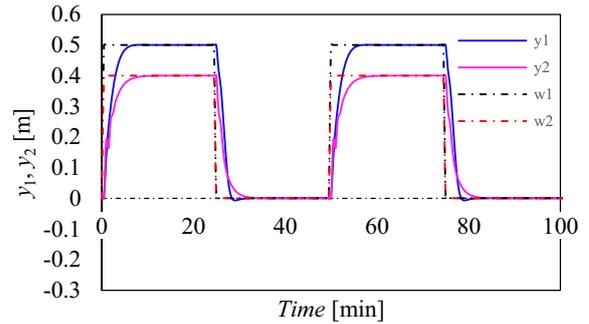


Figure 7: The reference and output signal of the gain scheduled control

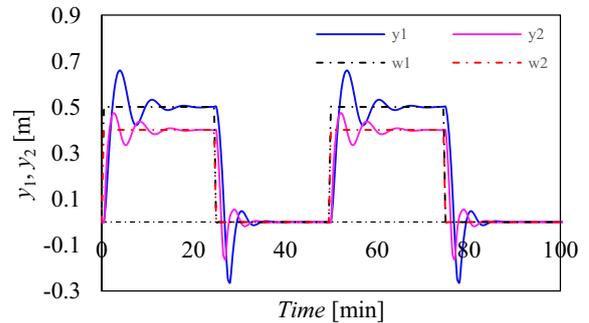


Figure 8: The reference and output signal of the fixed gain control

CONCLUSION

This paper addressed the problem of the gain scheduling procedure as well as the design of optimal control for a nonlinear multi-input multi-output system of two funnel liquid tanks in series. First, we have detailed studied the simplified model of the technological process. Based on the model, we have followed a general analytical framework for gain scheduling. We have also pointed out that selection of scheduling variable depends on particular characteristics of the system. This observation has critical importance and leads us to the conclusion that rule of scheduling on reference variable can be applied for other technological processes. The main advantage of this approach is that linear design methods can be applied to the linearized system at each operating point. Thanks to this feature, the presented procedure leaves room for many linear control methods. As our results show, presented integral control approach ensures robust regulation under all parameter perturbations. However the strength of the presented feedback lies in the optimal design of the gain matrix. In addition, we have demonstrated that a gain scheduled control system has the potential to respond rapidly changing operating conditions.

ACKNOWLEDGEMENT

This article was created with support of the Ministry of Education of the Czech Republic under grant IGA reg. n. IGA/FAI/2015/006.

REFERENCES

- Dostál, P.; V. Bobál; and F. Gazdoš. 2008. "Application of the polynomial method in adaptive control of a MIMO process". *Mediterranean Conference on Control and Automation - Conference Proceedings*, 131-136.
- Hairer, J. 1993. *Robust industrial control. Optimal design approach for polynomial systems*. Prentice Hall, Englewood E; S.P. Norsett; and G. Wanner. 1993. *Solving ordinary differential equations*. 2nd revised ed. Berlin: Springer.
- Jiang, J. 1994. "Optimal gain scheduling controllers for a diesel engine". *IEEE Control Systems Magazine*, 14(4), 42-48.
- Kaminer, I; A. M. Paswal; P. P. Khargonekar; and E. E. Coleman. 1995. "A velocity algorithm for the implementation of gain scheduled controllers". *Automatica*, 31, 1185-1191.
- Krhovják, A.; P. Dostál; and S. Talaš. 2014. "Multivariable adaptive control of two funnel liquid tanks in series". *Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014*, 273-278.
- Krhovják, A.; P. Dostál; S. Talaš. 2015; and L. Rušar. "Multivariable gain scheduled control of two funnel liquid tanks in series". in *Process Control (PC), 2015 20th International Conference on*, pp. 60-65.
- Lawrence, D. A. and W. J. Rugh. 1995. "Gain scheduling dynamic linear controllers for a nonlinear plant". *Automatica*, 31, 381-390.
- Shamma, J.S.; M. Athans. 1990. "Analysis of gain scheduled control for nonlinear plants. (1990) *IEEE Transactions on Automatic Control*, 35 (8), pp. 898-907.
- Shamma, J.S and M. Athans. 1992. "Gain scheduling: potential hazards and possible remedies". *IEEE Control Systems Magazine*, 12(3), 101-107.
- Shamma, J.S. and M. Athans. 1991. "Guaranteed properties of gain scheduled control of linear parameter-varying plants". *Automatica*, vol. 27, no. 4, 559-564.
- Rugh, W.J. 1991 "Analytical framework for gain scheduling". *IEEE Control Systems Magazine*, 11(1), pp. 79-84.
- Richardson, S.M. 1989. *Fluid mechanics*, New York, Hemisphere Pub. Corp.

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 interests 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.

LUKÁŠ RUŠAR studied at the Tomas Bata University in Zlín, Czech Republic, where he obtained his master degree in Automatic Control and Informatics in 2014. He now attends PhD. study in the Department of Process Control, Faculty of Applied Informatics of the Tomas Bata University in Zlín. His research interests focus on model predictive control. His e-mail address is rusar@fai.utb.cz.

BOND GRAPH MODEL OF A WATER HEAT EXCHANGER

Toufik Bentaleb, Minh Tu Pham, Damien Eberard, and Wilfrid Marquis-Favre
Univ Lyon, INSA, CNRS, AMPERE, F-69100, VILLEURBANNE, France
E-mail: name.surname@insa-lyon.fr

KEYWORDS

Bond graph; Multiport; Thermofluid systems; Plate heat exchanger

ABSTRACT

This paper presents a Bond Graph (BG) modelling approach to add and exploit on existing Modelica models some information on the energy structure of the systems. The developed models in the ThermosysPro library (Modelica-based) are already validated against the experimental data in previous works. A plate heat exchanger (PHE), which is equipment for nuclear power plants, is considered as a case study in this paper. Simulation results of the BG model for the counterflow PHE are compared with simulation results of the tested Modelica model. Comparisons show good agreement between both model results.

INTRODUCTION

Nowadays, the major problems of the numerical computation of mathematical models for complex processes are solved by using different commercial and open source software packages. The representation of the models in these languages is often based on model equations. The bond graph representation allows a physical structural analysis which is based on the system energy structure. This facilitates the exchange of models and simulation specifications. Bond graph is a graphical representation methodology for modelling multidisciplinary physical systems (Jardin et al. 2009).

Heat exchange is an important unit operation that contributes to efficiency and safety of many processes (nuclear power plants, steam generators, automotive, heat pumps, etc.). A plate heat exchanger is a type of heat exchanger that uses metal plates to transfer heat between two fluids. The plate heat exchanger was invented by Dr Richard Seligman in 1923 and revolutionised methods of indirect heating and cooling of fluids (Crepaco 1987). Plate heat exchangers are widely used in many other applications (food, oil, chemical and paper industries, HVAC, heat recovery, refrigeration, etc.) because of their small size and weight, their cleaning as well as their superior thermal performance compared to other types of heat exchangers (Guo et al. 2012).

The plate heat exchanger model is one of over 200 0D/1D models of components belonging to the ThermosysPro library. This open source library, developed by EDF R&D, is used to model energy systems and different types of power plants (nuclear, conventional, solar, etc.) (El Hefni 2014; El Hefni and Bouskela 2006; El Hefni et al. 2011, 2012; Deneux et al. 2013). The Modelica model is developed in Dymola. Modelica

representation leads to static analyses which are based on model equations, while BG representation permits a physical structural analysis which is based on the system's energy structure. The bond graph representation in this paper is built using the graphical editor MS1. MS1, an acronym of Modelling System One, is an interactive environment for modelling, simulation and analysis of non-linear dynamic systems (Jardin et al. 2008).

In the literature, bond graph modelling of heat exchangers is widespread (Shoureshi and Kevin 1983; Hubbard and Brewer 1981; Delgado and Thoma 1999). Due to difficulties in handling entropy and heat transfer rate, many efforts have been made to develop pseudo-bond graph representations of thermo-fluid transport and heat exchange (Karnopp and Azerbaijani 1981; Karnopp 1978, 1979; Ould Bouamama 2003). All these references mentioned above have different assumptions. For instance, in (Shoureshi and Kevin 1983) a temperature-entropy bond graph technique has been proposed based on three lump models to predict the reversal of flow. In this model, the authors have considered that the fluid domain is operated independently from the thermal domain. In (Karnopp 1978), pseudo bond graph strategies have been proposed with using the temperature and heat flow as effort and flow.

This paper uses pseudo bond graph method for heat/mass transfer modelling. Furthermore, multi-port C and multi-port R elements have been used. This method is based on finite volume approach considering the thermal and fluid bonds. First, fundamental theory of thermofluid is given. Then, the plate heat exchanger models are explained: the Modelica model and the BG model. In the section after that, simulation results are discussed. The last section contains conclusion and future research paths.

THERMOFLUID SYSTEM

Thermofluid or thermal fluid sciences involve the study of the thermodynamics, fluid mechanics, heat and mass transfer in complex engineering systems. In the open system case, the energy and mass equations for a thermodynamic system are formulated as (see nomenclature page 7)

$$\frac{dE_s}{dt} = \dot{Q}_{in} + \dot{W}_{in} + \dot{E}_i - \dot{E}_e \quad (1)$$

where (dE_s/dt) is the rate of increase in energy within the system, \dot{Q}_{in} is the rate at which heat enters the system, \dot{W}_{in} is the rate at which work enters the system, \dot{E}_i is the rate at which energy is brought in by the mass entering the system, and \dot{E}_e is the rate at which energy is removed by the mass leaving the system.

$$\frac{dm_s}{dt} = \dot{m}_i - \dot{m}_e \quad (2)$$

where (dm_s/dt) represents the rate of increase in mass within the system, and m_i and m_e represent the respective rates at which mass entering and leaving the system.

In many thermal application, the reduced heat equation is used

$$\dot{Q} = \frac{dQ}{dt} = mC_p \frac{dT}{dt} = KA(\Delta T), \quad (3)$$

where \dot{Q} is the heat-flow-rate (named just heat rate), C_p is the specific heat capacity, m is the mass flow rate, the global heat transfer coefficient K (associated to a bounding area A and the average temperature jump ΔT between the system and the surroundings). More details about the above equations can be seen in reference (Martínez 1992).

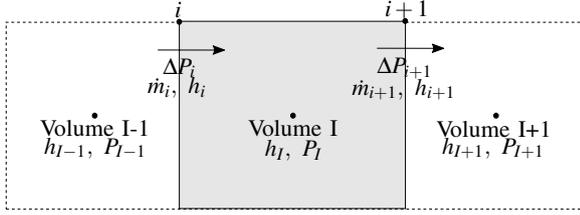


Figure 1: Staggered finite volume scheme

The cooling water heat exchanger used in this study is an equipment for nuclear power plants. Two main approach for the dynamic modelling of the heat exchanger are the moving boundaries (MB) and the discretized models, known as finite-volume models (FVM) (Bendapudi et al. 2004; Desideri et al. 2015). The moving boundary method is useful for developing feedback controllers, in this approach the heat exchanger is divided into zones based on the fluid phase in each region and the location of the boundary between regions vary in time according to the current conditions. In finite-volume models the 1D flow is subdivided into several equal control volumes as shown in Figure 1.

The modelling technique used in this paper is based on finite volumes approach. A pictorial representation of the discretized counterflow heat exchanger is shown in Figure 2.

PLATE HEAT EXCHANGER MODELS

The plate heat exchanger is the component that transforms heat (thermal energy) from one fluid to another. Plate heat exchangers have a high heat transfer rate compared to other types of heat exchangers due to their large surface area.

Modelling of Water/Water Heat Exchangers in Modelica

The dynamic water/water heat exchanger component used belongs to the ThermoSysPro library. The core model of the heat exchanger was written in Modelica and simulated with the Dymola simulation environment. Figure 3 shows the schematic of the heat exchanger model in Dymola. This model has two parts: the upper part for hot water and the other part for cold water,

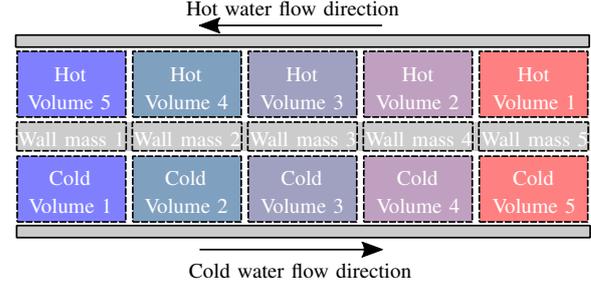


Figure 2: Schematic diagram of a typical discretized counterflow plate heat exchanger

which are quite similar. The inlets and outlets of the heat exchanger block are shown by the blue and red rectangles, respectively. The red and blue arrows show the cold and hot waters of the two parts respectively.

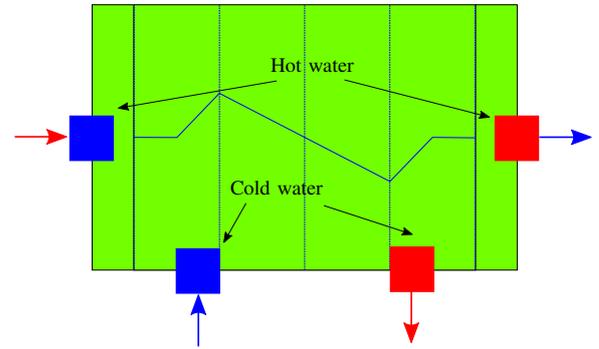


Figure 3: Dymola layout of the heat exchanger

In this model, the rate of mass accumulation within the volume does not incorporate any dynamic effects. This means that the entering mass flow rate is exactly equal to the leaving one i.e. the steady balance for all volumes of the heat exchanger, yields

$$\dot{m}_{b,i-1} - \dot{m}_{b,i} = 0 \quad (4)$$

where \dot{m} is the mass flow rate. Throughout the paper, the subscript b means the hot part when $(b \leftarrow h)$ or the cold part when $(b \leftarrow c)$.

To simplify the model, the mass flow rate is considered positive in both parts i.e. $\dot{m}_{b,i} > 0$, and the pressure between each two volumes is defined as

$$P_{b,i+1} = P_{b,i} - \Delta P_{b,i}/N \quad (5)$$

where N is number of segments, $\Delta P_{b,i}$ is the pressure drop.

The pressure drop ($\Delta P_{b,i}$), which has direct relationship to the size of the plate heat exchanger, is defined by

$$\Delta P_{b,i} = k_{b,i} \cdot Nu_{b,i}^{-a} \cdot qu_{b,i}^2 + 104.97 \cdot Nu_{b,i}^{-0.25} \quad (6)$$

where $a = 0.097$ and $k_{b,i}$, correlation for the heat transfer Nu (is called also Nusselt number), and pressure drop qu characteristics, are defined as

$$\begin{cases} Nu_{b,i} = \dot{m}_{b,i}/(M \cdot \mu_{b,i}) \\ qu_{b,i} = \dot{m}_{b,i}/M \\ k_{b,i} = 14423.2 \left[1472.47 + \frac{1.54(M-1)}{2} \right] \frac{c_{1,b}}{\rho_{b,i}} \end{cases} \quad (7)$$

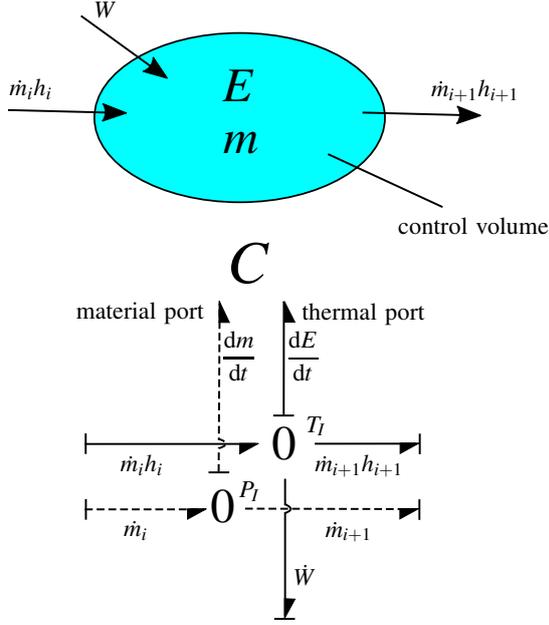


Figure 5: Control volume and its corresponding bond graph of the thermodynamic accumulator

$$\text{Thermal Bond } T_{b,I} = T_{b,I}(P_{b,I}, h_{b,I}) \quad (16)$$

where the temperature table $T_{b,I}(P_{b,I}, h_{b,I})$, thermodynamic property of the water, is well detailed in reference (Wagner and Kretschmar 2008) and the ThermoSysPro model. The function $P_{b,I}(\rho_{b,I}, h_{b,I})$ is calculated by an iterative resolution given in Algorithm 1 described in the appendix section. The algorithm provides an easiest way to obtain the pressure, as a function of a given density and specific enthalpy. To obtain the density the dynamic continuity equation, in which the total stored mass equals the net integrated mass flow rate, is used

$$\frac{dm_{b,I}}{dt} = \dot{m}_{b,i} - \dot{m}_{b,i+1} \quad (17)$$

where $m_{b,I}$ is the mass, $\dot{m}_{b,i}$ and $\dot{m}_{b,i+1}$ are respectively the control volume entering and leaving mass flow rates, as shown in Figure 5 by the dashed bonds.

To convert Equation (17) into a more useful form to obtain the density in the control volume, the following relationship is used

$$\rho_{b,I} = \frac{m_{b,I}}{V_{b,I}} \quad (18)$$

where $V_{b,I}$ is the volume of the control volume.

Assuming a C-element with a constant volume, which leads to

$$\frac{d\rho_{b,I}}{dt} = \frac{\dot{m}_{b,i} - \dot{m}_{b,i+1}}{V} \quad (19)$$

In Equations (16) and (15), the specific enthalpy $h_{b,I}$ is calculated from the following one-dimensional energy equation

$$V_{b,I} \cdot \rho_{b,I} \cdot \frac{dh_{b,I}}{dt} = \pm (h_{b,i} \cdot \dot{m}_{b,i} - h_{b,i+1} \cdot \dot{m}_{b,i+1} - \dot{W}_I) \quad (20)$$

- **Multi-port R-element,**

The Multi-port R-element is used for a thermodynamic resistance and in its pseudo-bond graph form. The structure of the multiport pseudo-bond graph model R-element is defined in Figure 6.

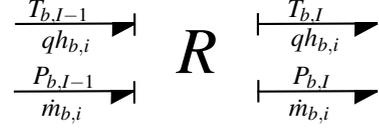


Figure 6: The pseudo bond graph of the R-field of the heat exchanger

The relationships between the efforts (temperatures and pressures) and flows (mass flow rate and specific enthalpy flow rate) in R-elements are given by

$$\begin{cases} \dot{m}_{b,i} = \dot{m}_{b,i}(T_{b,I-1}, T_{b,I}, P_{b,I-1}, P_{b,I}), \\ qh_{b,i} = qh_{b,i}(T_{b,I-1}, T_{b,I}, P_{b,I-1}, P_{b,I}). \end{cases} \quad (21)$$

where the specific enthalpy flow rate $qh_{b,i}$ is represents the quantity $\dot{m}_{b,i}h_{b,i}$ shown in the Figure 6.

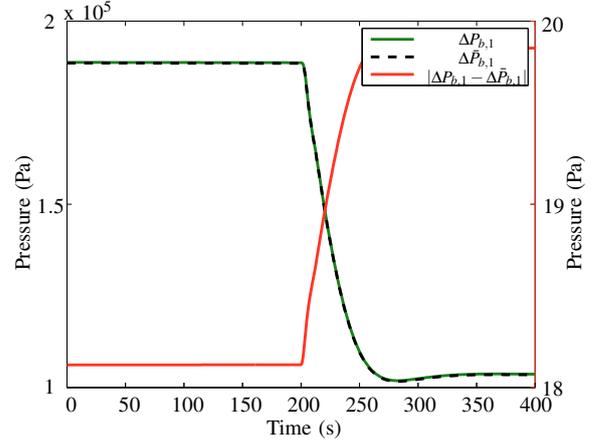


Figure 7: Comparison between the pressure difference

To calculate the mass flow rate $\dot{m}_{b,i}$ as a function of $\Delta P_{b,i}$ using the Equation (6) is quite difficult. Hence, to obtain an approximate solution, we ignore the second term $104.97 \cdot (Nu_{b,i})^{-0.25}$ in the Equation (6) because its effect on $\Delta P_{b,i}$ is negligible (see Figure 7). The pressure drop can be written as

$$\begin{aligned} \Delta \bar{P}_{b,i} &= k_{b,i} \cdot Nu_{b,i}^{-a} \cdot qu_{b,i}^2 \\ &= k_{b,i} \cdot \left(\frac{\dot{m}_{b,i}}{\mu_{b,i} \cdot M} \right)^{-a} \cdot \left(\frac{\dot{m}_{b,i}}{M} \right)^2 \\ &= \bar{k}_{b,i} \cdot \dot{m}_{b,i}^{2-a} \end{aligned} \quad (22)$$

where $\bar{k}_{b,i} = k_{b,i} \cdot \mu_{b,i}^a \cdot M^{(a-2)}$. Thus, the mass flow rate $\dot{m}_{b,i}$ can be calculated by using the following formula

$$\dot{m}_{b,i} = \exp \left(\frac{\ln(\Delta \bar{P}_{b,i}) - \ln(\bar{k}_{b,i})}{2-a} \right) \quad (23)$$

Figure 8 shows the comparison between both mass flow rates of the hot water in Dymola and MS1, in which calculated by Equation (23). The blue curve represents the mass flow rate obtained by the ThermoSysPro model,

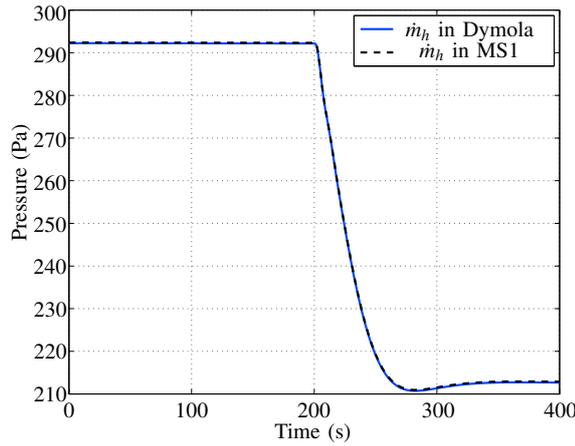


Figure 8: Comparison between both mass flow rates in Dymola and MS1

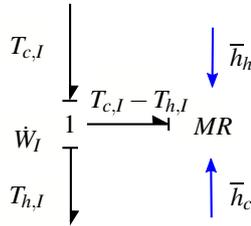


Figure 9: Modulated 1-port R-element and 1-junction

and dashed black curve represents the mass flow rate obtained by BG model. As a result, both curves are similar because the pressure drop obtain by Equation (22) is quite equal to that obtained by Equation (6).

- **Modulated 1-port R-element,**

The rate of heat transfer in each segment occurs between two C-fields at higher and lower temperatures. The heat flow between C-fields (from higher temperature to lower temperature) is proportional to the difference of temperatures, and is given by

$$\dot{W}_I = K_I \cdot \Delta S \cdot (T_{h,I} - T_{c,I}) \quad (24)$$

where K_I and ΔS are given in Eq (10). K_I is a function of the convection heat transfer coefficient \bar{h}_b , the latter is calculated in R-Element and materialized by the blue arrows in Figure 9.

In Figure 9, the 1-junction means that the rate of heat transfer \dot{W} (flow) through all connected bonds is the same, and that the temperatures (efforts) sum to zero. The sign of each temperature is related to the power direction (i.e. direction of the half arrow) of the bond.

In the bond graph model, the calculation of some thermodynamic properties of water is different compared to Equation (12). Here, the thermodynamic properties of water are given in terms of the pressure and the

temperature as following

$$\begin{cases} \rho_{b,i} = \rho_{b,i}(P_{b,i}, T_{b,i}) \\ h_{b,i} = h_{b,i}(P_{b,i}, T_{b,i}) \\ \lambda_{b,i} = \lambda_{b,i}(\rho_{b,i}, T_{b,i}) \\ \mu_{b,i} = \mu_{b,i}(\rho_{b,i}, T_{b,i}) \\ Cp_{b,i} = Cp_{b,i}(P_{b,i}, T_{b,i}) \end{cases} \quad (25)$$

where, the pressure $P_{b,i}$ and the temperature $T_{b,i}$ at the multi-port R-element are given by

$$\begin{cases} P_{b,i} = \frac{P_{b,I-1} + P_{b,I}}{2} \\ T_{b,i} = \frac{T_{b,I-1} + T_{b,I}}{2} \end{cases} \quad (26)$$

The functions in Equation (25) of the water properties are based on the Industrial Formulation IAPWS-IF97 which consists of a set of equations for different water regions (more details are given in (Wagner and Kretzschmar 2008)).

RESULTS AND DISCUSSION

The bond graph model of the exchanger has been validated by simulations. Figure 10 shows the bond graph model of the PHE. The PHE model consists of total ($N = 5$) numbers of plates, each layer being represented as a small heat exchanger (shown in Figure 4). The inputs of bond graph model are the pressure and the temperature, while the outputs are the mass flow rate and enthalpy flow rate for hot and cold parts.

The Modelica and the BG models were run for 400 seconds of simulation time. Figure 11 shows the pressures, where the black curves represent the pressure at the boundaries, the blue curves represent the pressure at each volume in the ThermoSysPro model, and the dashed green curves represent the pressure at each volume (C-element) in the BG model. Figure 12 shows the error between obtained pressures (Dymola and MS1) at first volume. From the comparison of the simulation results, the conclusion that can be made is that the both models have similar dynamic behavior.

CONCLUSION

In this paper, a multi-port pseudo BG model of a plate heat exchanger system has been presented. The model can be used in transient system simulations and can be extended to cover other heat exchanger types. The comparison of the simulation results of bond graph model with the Modelica model indicates that the model predicts the dynamic behavior of the heat exchanger well.

Future developments will include developing bond graph models of centrifugal pump, regulating valve, feeding on-off valve, and pipes for nuclear power plants. The main objective is to study the observability, initial conditions, structural inversibility by physical structural analysis to improve systems diagnosis and operation.

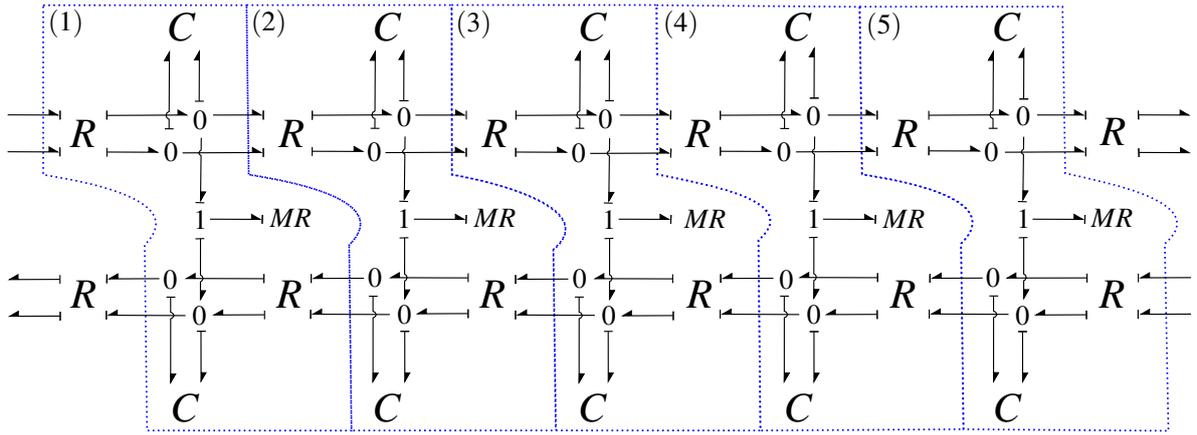


Figure 10: Pseudo bond graph of the plate heat exchanger (counterflow)

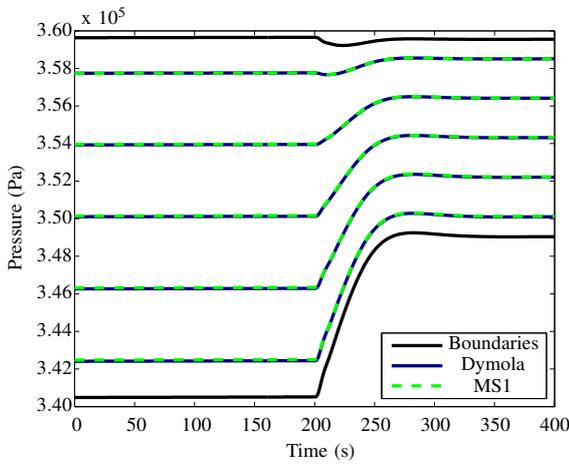


Figure 11: Comparison of bond graph model with Dymola model in hot flow

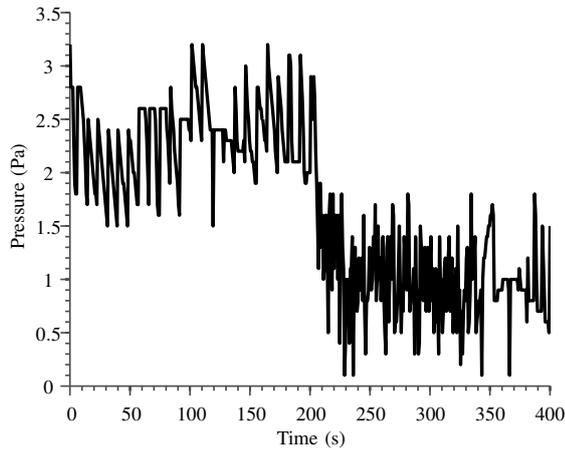


Figure 12: Error between obtained pressures (Dymola and MS1) at first volume

APPENDIX

The following algorithm has been developed for pressure calculation in the C-accumulator of the heat exchanger. Based on the density function it was developed to determine water pressure in terms of specific enthalpy and density. The structure of the algorithm is shown in Algorithm 1. This is a very fast and

accurate method in all regions except solid region of the water. This algorithm is used because, as far as we know (Wagner and Kretzschmar 2008), there are no tables to calculate the water pressure directly in terms of specific enthalpy and density. In the solid region, the calculation of the water pressure is based on incompressibility consideration.

```

Data:  $\rho, h;$ 
 $p_{min} = 0.00611657; p_{max} = 1000;$ 
 $p_s = 100; \rho_s = \rho_s(p_s, h);$ 
while  $|\rho - \rho_s| > 1E-7$  do
   $\rho_s = \rho_s(p_s, h);$ 
  if  $\rho_s \geq \rho$  then
     $p_{max} = p_s;$ 
  else
     $p_{min} = p_s;$ 
  end
   $p_s = (p_{min} + p_{max})/2;$ 
end

```

Algorithm 1: Calculation of the water pressure using the density function

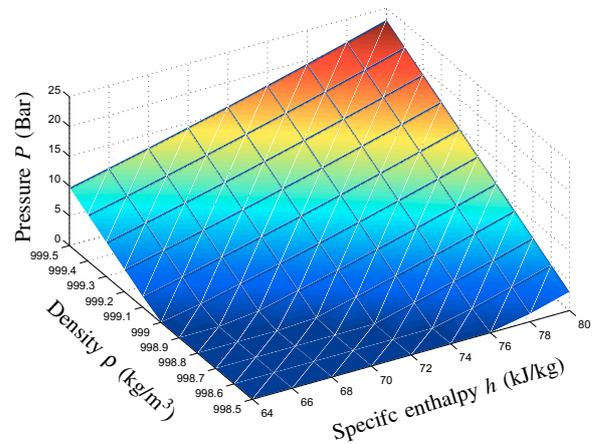


Figure 13: Pressure as a function of h and ρ

Figure 13 shows how the thermodynamic propriety of the water pressure is very sensitive to small changes in density of the water.

ACKNOWLEDGEMENTS

The research described in this paper is partly supported by DGE within the ongoing ITEA2 (www.itea2.org) project 11004 "MODRIO" (Model Driven Physical Systems Operation). The authors would like to thank Audrey Jardin and Daniel Bouskela (EDF R&D) for their help and valuable comments on this paper.

NOMENCLATURE

λ_m	Metal thermal conductivity	P	Pressure
\dot{m}	Mass flow rate	ΔP	Pressure drop
h	Specific enthalpy	ρ	Density
ΔS	Heat transfer surface	C_p	Heat capacity
L	Heat exchanger length	T	Temperature
n	Number of plates	Pr	Prandtl number
μ	Dynamic viscosity	\dot{Q}	Heat flow rate
e_m	Metal wall thickness	\dot{W}	Thermal power
K_f	Heat transfer coefficient	V	Volume
N	Number of segments	Nu	Nusselt number

Chosen abbreviations

PHE	Plate Heat Exchanger
BG	Bond Graph
FVM	Finite-Volume Method
MB	Moving Boundaries

Subscripts

I	in the volume I	$b = c$	Cold side
i	entering the volume I	$b = h$	Hot side

REFERENCES

- Bendapudi, S., Braun, J. E. and Groll, E. A. (2004). "Dynamic Modeling of Shell-and-Tube Heat-Exchangers: Moving Boundary vs. Finite Volume", *International Refrigeration and Air Conditioning Conference*, Purdue.
- Cracow, T. S. (2011). "Experimental Prediction of Heat Transfer Correlations in Heat Exchangers", *University of Technology*, Poland.
- Crepaco, A. (1987). "Heat Transfer Handbook: Design and Application of Paraflo Plate Heat Exchangers", *APV Crepaco, Incorporated*.
- Delgado, M. and Thoma, J. (1999). "Bond Graph Modeling and Simulation of a Water Cooling System for a Moulding Plastic Plant", *Systems Analysis Modeling Simulation*, Vol. 36, pp. 153–171.
- Deneux, O., El Hafni, B., Péchiné, B., Di Penta, E., Antonucci, G. and Nuccio, P. (2013). "Establishment of a Model for a Combined Heat and Power Plant With ThermoSysPro Library", *Procedia Computer Science*, Vol. 19, pp. 746–753.
- Desideri, A., Wronski, J., Dechesne, B., van den Broek, M., Gusev, S., Quoilin, S. and Lemort, V. (2015). "Comparison of Moving Boundary and Finite-Volume Heat Exchangers Models in the Modelica Language", *3rd International Seminar on ORC Power Systems*, Brussels, Belgium.
- El Hefni, B. (2014). "Dynamic Modeling of Concentrated Solar Power Plants With the ThermoSysPro Library (Parabolic Trough Collectors, Fresnel Reflector and Solar-Hybrid)", *Energy Procedia*, Vol. 49, pp. 1127–1137.
- El Hefni, B. and Bouskela, D. (2006). "Modeling of a Water/Steam Cycle of the Combined Cycle Power Plant "Rio Bravo 2" with Modelica", *Modelica conference proceedings*.
- El Hefni, B., Bouskela, D. and Lebreton, G. (2011). "Dynamic Modelling of a Combined Cycle Power Plant With ThermoSysPro", *Proceedings of the 9th International Modelica Conference*, Munich, Germany.
- El Hefni, B., Bouskela, D. and Lebreton, G. (2012). "Dynamic Modelling of a Condenser/Water Heat With ThermoSysPro", *Proceedings of the 8th International Modelica Conference*, Vol. 63, Linköping, Linköping University Electronic Press, pp. 365–375.
- Guo, C., Du, W. and Cheng, L. (2012). "Characteristics of Heat Transfer and Resistance of Double Chevron Plate Heat Exchanges With Different Corrugation Pitch", *Advances in Intelligent and Soft Computing*, Vol. 143, pp. 169–174.
- Hubbard, M. and Brewer, J. (1981). "Pseudo Bond Graphs of Circulating Fluids with Application to Solar Heating Design", *Journal of the Franklin Institute*, Vol. 311, pp. 339–354.
- Jardin, A., Marquis-Favre, W. and Thomasset, D. (2009). "Bond Graph Sizing of Mechatronic Systems: Coupling of Inverse Modelling with Dynamic Optimization", *I. Troch, F. Breitenacker. MATHMOD*, Vienne, Austria.
- Jardin, A., Marquis-Favre, W., Thomasset, D., Guillemard, F. and Lorenz, F. (2008). "Study of a Sizing Methodology and a Modelica Code Generator for the Bond Graph Tool MS1", *Proceedings of the 6th International Modelica Conference*, Bielefeld, Allemagne, pp. 125–134.
- Karnopp, D. (1978). "Pseudo Bond Graphs for Thermal Energy Transport", *Journal of Dynamic Systems, Measurement and Control*, Vol. 100, pp. 165–169.
- Karnopp, D. (1979). "State Variables and Pseudo-Bond Graphs for Compressible Thermo-Fluid Systems", *Journal of Dynamic Systems, Measurement and Control*, Vol. 101, pp. 201–204.
- Karnopp, D. and Azerbaijani, S. (1981). "Pseudo Bond Graphs for Generalised Compartmental Models in Engineering and Physiology", *Journal of the Franklin Institute*, Vol. 312, pp. 95–108.
- Karnopp, D. C., Margolis, D. L. and Rosenberg, R. C. (1990). "System Dynamics: A Unified Approach", *John Wiley & Sons, Inc.*.
- Karnopp, D., D.L., M. and Rosenberg, R. (2012). "System Dynamics: Modeling, Simulation, and Control of Mechatronic Systems, Fifth Edition", *John Wiley & Sons, Inc.*, Hoboken, NJ, USA.
- Martínez, I. (1992). "Termodinámica Básica y Aplicada", *Ed. Dossat*.
- Ould Bouamama, B. (2003). "Bond Graph Approach as Analysis Tool in Thermo-Fluid Model Library Conception", *Journal of the Franklin Institute*, Vol. 340, pp. 1–23.
- Shoureshi, R. and Kevin, M. (1983). "Analytical and Experimental Investigation of Flow-Reversible Heat Exchangers Using Temperature-Entropy Bond Graphs", *American Control Conference*, pp. 1299–1304.
- Wagner, W. and Kretschmar, H.-J. (2008). "International Steam Tables - Properties of Water and Steam Based on the Industrial Formulation IAPWS-IF97", Springer-Verlag Berlin Heidelberg.

MODELLING A PCT40 HEAT EXCHANGER FOR CONTROL PURPOSES

Frantisek Gazdos and Daniel Macek
Faculty of Applied Informatics
Tomas Bata University in Zlin
Nam. T. G. Masaryka 5555, 760 01 Zlin, Czech Republic
E-mail: gazdos@fai.utb.cz

KEYWORDS

Modelling, Simulation, Heat exchanger, PCT40.

ABSTRACT

This paper presents a simulation model of a PCT40 heat exchanger. It starts with a motivation, followed by description of the system and of the whole modelling process. The mathematical model includes simplified description of heating and cooling systems together with a more detailed description of a proportioning solenoid valve. It is formed using combination of both analytical and empirical approaches and the resultant simulation model is compared to real-time measurements in terms of both static and dynamic properties. The overall results indicate that although the model is relatively simple it can be used for control-oriented simulation experiments, which was the purpose of this work and saves time during experiments on this system.

INTRODUCTION

Modelling and simulation courses are essential parts of education process in the field of control engineering. Students – prospective control experts have to be able to model a given system to be controlled in order to design a model-based controller. Prior to real-time implementation of the proposed controller it is also common nowadays to perform simulation testing to ensure the proposed control system is effective and safe at the same time. In order to prepare students for industrial practice properly the courses always include a form of laboratory exercises where students test their knowledge of control field on laboratory-scale industrial processes. These labs are usually included in the follow-up studies and students face real control problems here – from system identification, controller design to real-time implementation and tuning on different plants related to industrial practice. Similarly, during studies of Automatic Control & Informatics Master's degree course at the Faculty of Applied Informatics, Tomas Bata University in Zlin (FAI TBU in Zlin 2016), students have to complete the course “Real Process Control”. Here students work in the Department of Process Control laboratory with various laboratory-scale processes to prove their understanding of control engineering. In this lab there is a wide range of processes from different producers to ensure students test many different types of systems. For instance,

TecEquipment models (TecEquipment 2016) include CE107 Engine Speed Control Apparatus, CE108 Coupled Drives Apparatus, CE120 Controller, CE150 Helicopter Model, CE151 Ball and Plate Apparatus and CE152 Magnetic Levitation Model. AMIRA models (Amira 2016) consists of DTS200 Three-Tank-System, DR300 Speed Control with Variable Load and PS600 Inverted Pendulum. Then there are also other models from Feedback (Feedback 2016): 33-007-PCI Twin Rotor MIMO System, from Leybold (LD Didactic 2016): T 8.2.1.4 Gas Flow Control and from Armfield (Armfield 2016): PCT40 Multifunction Process Control Teaching System. While some of the processes are fast, e.g. the magnetic levitation model, coupled drives apparatus or speed control system – experiments with these models are relatively quick with responses in seconds, others can be really slow with responses in tens of minutes or even hours. These include e.g. the three-tank-system or a heat-exchanger from the multifunction process control teaching unit. In this case students have to wait even hours to complete their experiment. Therefore it is very useful for them to have simulation models of these systems that capture their main properties so that students or other faculty staff can test their control algorithms by simulation means, prior to real-time implementation. Simulation experiments with these models are much faster with responses in seconds which leads to more effective work. For instance, a simulation model of the three-tank-system was developed and presented in (Chalupa et al. 2012). Modelling of the PCT40 Heat Exchanger has been the subject of student's Master's thesis (Macek 2015) and it is described briefly in this contribution, including also main results. The modelling principles employed here follow the basic guidelines presented in the modelling classics, e.g. (Wellstead 1979; Ljung and Torkel 1994; Severance 2001). A simplified, first-principles mathematical model is derived analytically with parameters further tuned using real-time experiments.

The paper is structured as follows: after this introductory section it continues by a description of the PCT40 heat exchanger including also experimental conditions and main variables. The main body includes modelling of the basic process parts – heating/cooling systems and proportioning solenoid valve (PSV). Last section compares the developed model to real-time data and the paper concludes summarizing main results and suggesting future possible directions of development.

PCT40 HEAT EXCHANGER

The heat exchanger modelled in this paper is a part of the PCT40 multifunction process control teaching system (Armfield 2016). This unit is designed for use in teaching a wide range of control methods and to demonstrate a variety of process control loops. Processes such as level, temperature, flow or pressure control can be easily implemented using the provided interface, multifunction I/O card and suitable software. More advanced aspects of control can be addressed by adding optional extras to the basic system, such as fluid property control (conductivity and pH probe), remote set points or dual loops (Armfield 2005). Illustrative overview of the unit is presented in Fig. 1. The heat exchanger is located on the right side and it is further depicted schematically in more detail in Fig. 2.



Figure 1: Basic Process Control Unit PCT40 with Process Vessel Accessory

Heat Exchanger Description

It consists of a small clear acrylic vessel that incorporates an electrical element for heating water. A thermostat and level detector are incorporated in the vessel to prevent the heater from operating if the water is too hot ($>80\text{ }^{\circ}\text{C}$) or the level in the vessel is too low. The vessel lid provides support for the stainless steel heating/cooling coil. Fittings at the inlet and outlet of the coil accommodate thermocouple-type temperature sensors and allow connection to the water supply. The lid also accommodates adjustable glands for a variable-height thermocouple sensor, a thermometer pocket and a temperature switch (thermostat) to be inserted into the vessel for the purpose of calibration. The main water inlet is connected to a pressure regulating valve with integral filter. The flow of water through the equipment can be varied by adjusting the setting of the regulator. A turbine type flow meter is fitted in series with the main water inlet to allow inlet flow rate measurement. It has a range from 0.2 to 1.5 litres/minute approximately. A proportioning solenoid valve (PSV) is also located near the main water inlet to enable continuous regulation of the inlet flow rate. The heater power can be controlled by SSR (Solid State Relay) drive in on/off sense only

with nominal power 2 kW. Three temperature sensors – k-type thermocouple probes ($0\text{--}200\text{ }^{\circ}\text{C}$) are incorporated within the vessel. The first one is used to measure the fluid temperature inside the vessel while the two remaining located at the inlet/outlet of the heating/cooling coil are used to measure the fluid temperature as it enters and leaves the coil (Armfield 2005).

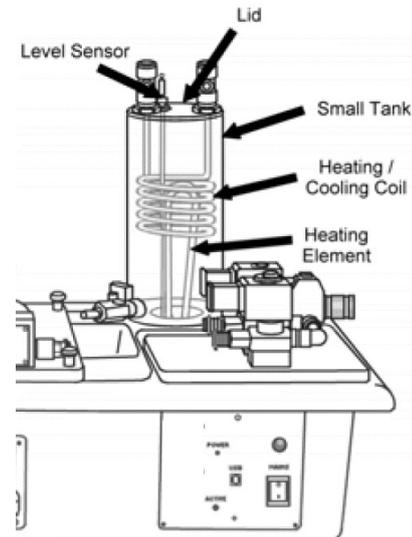


Figure 2: Heat Exchanger Schema

Experimental Conditions

Within the “Real Process Control” course during experiments on this unit in the Department of Process Control laboratory, Faculty of Applied Informatics, Tomas Bata University in Zlin, students have a special set-up of this system. First the vessel is filled with water to a certain level to ensure that both heating element and heating/cooling coil are under water. The level stays the same during the experiments. The coil is used for cooling only and so it is connected to cold water inlet using the pressure regulating valve, proportioning solenoid valve and the flow meter. The inlet flow rate is set to 1500 ml/min approx. using the pressure regulating valve and then it can be controlled continuously using the PSV valve via the I/O card and MATLAB/Simulink software. Then students perform various experiments including e.g. measurements of the static properties of the PSV valve, pulse-width modulation (PWM) of the heater power for temperature control, identification of heating/temperature and cooling/temperature systems, temperature control using heating and cooling, etc.

Main Variables for Modelling and Control

From the control theory point of view the main usual output variable to be controlled is the temperature inside the vessel $T(t)$ [$^{\circ}\text{C}$], measured via the k-type thermocouple sensor. Control inputs (manipulated variables) used for the temperature control are usually heater power $P(t)$ [%] (regulated using PWM) and water flow rate through the cooling coil $q_c(t)$ [l/min]

(measured by the turbine type flow meter and controlled using PSV). Therefore it can be generally seen as a MIMO system with 2 inputs and 1 output. For modelling purposes the main state-variables were selected as: water level in the vessel $h(t)$, temperature inside $T(t)$ and (output) temperature of the cooling water $T_C(t)$.

MATHEMATICAL MODEL

The modelling goal was to develop a relatively simple mathematical model for control purposes that captures main static and dynamic properties of the system rather than deriving a comprehensive mathematical model. Therefore the heat-exchanger is modelled as a lumped-parameters system using a set of ordinary differential equations (ODEs) only. The presented model consists of two main parts – model of the PSV valve and model of the heat exchanger that can be also divided into two interconnected systems – heating and cooling models.

Model of Proportioning Solenoid Valve

This regulating valve has specific static and dynamic properties which has been measured and identified experimentally and then modelled mathematically. First, a static characteristics has been measured – during repeated experiments the valve was opening gradually and the resultant flow rate was recorded, and the same for the valve closing. Repeated experiments were averaged. The results are presented in Fig. 3 below.

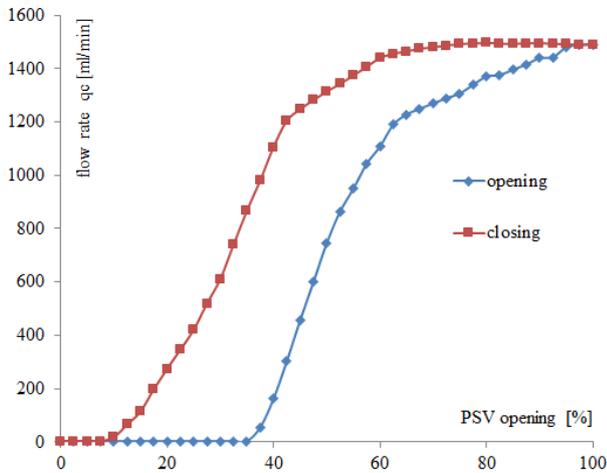


Figure 3: PSV – Static Properties

From the graph it is clear that the valve has a hysteresis and that suitable (nearly linear) area for control is between 40 – 60 [%] for opening and 20 – 40 [%] for closing approximately. The opening and closing curves were approximated using the polynomial regression of 5th degree and consequently the resultant equations are of the form (1) for opening and (2) for closing.

$$q_{c/o} = -1.754 \times 10^{-5} \cdot o^5 + 0.0061 \cdot o^4 - 0.83 \cdot o^3 + 53.23 \cdot o^2 - 1577.6 \cdot o + 17234 \quad (1)$$

$$q_{c/c} = -1.43 \times 10^{-6} \cdot o^5 + 5.21 \times 10^{-4} \cdot o^4 - 0.07 \cdot o^3 + 3.43 \cdot o^2 - 34.91 \cdot o + 62.31, \quad (2)$$

where the variable o [%] describes the opening/closing rate. The approximation is presented in Fig. 4 and shows relatively good fit of both curves to the measured data.

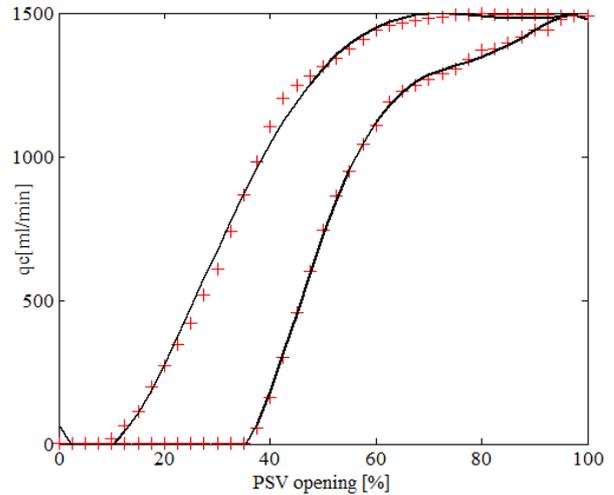


Figure 4: PSV – Approximation of Opening/Closing

Dynamical properties were measured as step-responses in different operating points for both opening and closing. As the responses were relatively similar the results were averaged, normed and identified as a first-order system with a time-constant $T = 0.33$ [s] approximately, i.e. with a transfer function:

$$G(s) = \frac{1}{0.33s + 1}. \quad (3)$$

Comparison of this model for both opening and closing responses are given in the following graphs, Fig.5 and Fig.6. The results confirm good fit of this simple model.

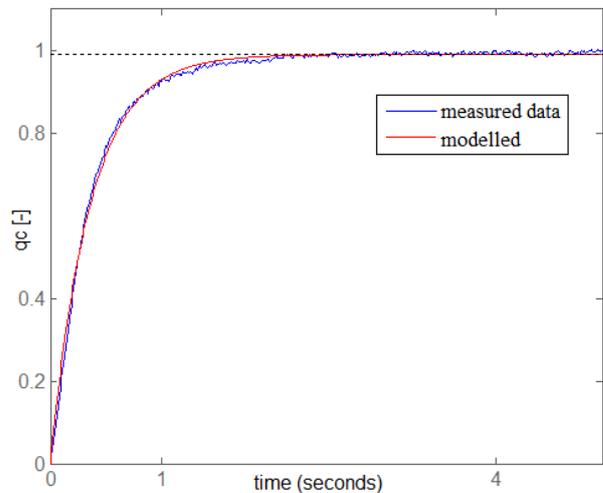


Figure 5: PSV – Dynamical Approximation of Opening

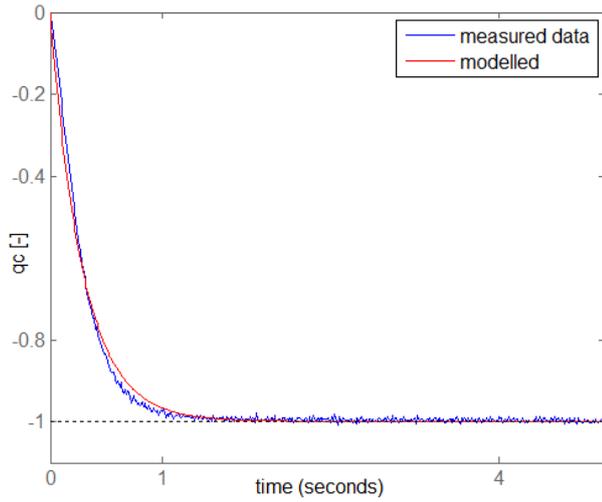


Figure 6: PSV – Dynamical Approximation of Closing

Both static and dynamic parts of this model were combined in the MATLAB/Simulink to form a simplified simulation model of the PSV. Input into the model is the value of valve opening in [%] and the model gives flow rate as the output variable in [ml/min]. There are also several auxiliary functions and blocks to detect opening/closing and for the hysteresis.

Model of Heat Exchanger

As it has been explained before, the mathematical model of the heat exchanger can be divided into two main interconnected parts – the water filled vessel with a heating element and the coil with cooling water. Schematic picture of the heat exchanger is presented in the following figure (Fig. 7). In the picture, q_v describes the input flow rate into the vessel of temperature T_v , output flow rate is denoted as q with temperature T . Water volume V in the tank is function of water level h and the tank diameter D . Heater power is denoted as P and its efficiency as η . Heat transfer coefficients are described using symbols α (heated water ↔ surroundings with temperature T_o) and α_c (heated water ↔ cooling water). Input/output flow rate of cooling water is denoted as q_c with input and output temperatures T_{cv} , T_c , respectively. Cooling water volume describes variable V_c . As it has been explained earlier in the paper the main state-variables for modelling were selected as: water level in the tank $h(t)$, temperature inside $T(t)$ and (output) temperature of the cooling water $T_c(t)$. Consequently the model is described by 3 basic equations derived using mass and energy balances of the whole system. The following simplified assumptions were taken into account for the modelling purposes: ideal mixing of both heated and cooling liquid, heat capacity of the vessel and cooling coil walls is neglected and all the following variables remain constant during the experiment: heat transfer coefficients α and α_c , densities of both fluids ρ and ρ_c and the same for their heat capacities c_p , c_{pc} .

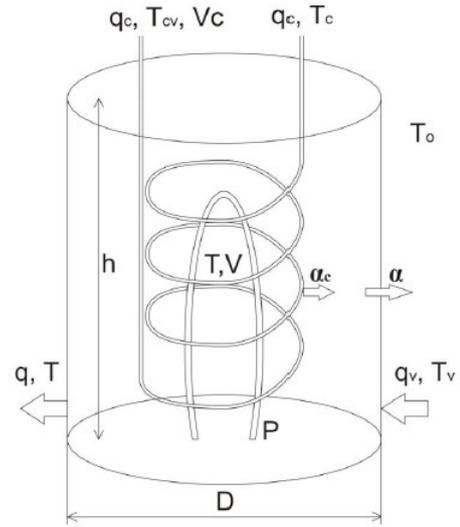


Figure 7: Heat Exchanger Main Variables

Material balance of heated water can be described by:

$$q_v(t) = q(t) + \frac{dV(t)}{dt}, \quad (4)$$

where the output flow rate can be modelled simply as:

$$q(t) = k\sqrt{h(t)} \quad (5)$$

for some constant k . The volume can be further described using the water level and cross-section area F as $V = Fh(t)$ while ideally $F = \pi D^2/4$. In fact, the cross-section area D is smaller as there is a heating element and cooling coil inside the vessel. Therefore it was calculated experimentally for several volumes from the equation above and averaged. The result is $F = 0.00705 \text{ m}^2$ approx. (while ideally it can be easily calculated using its diameter $D = 0.1 \text{ m}$ as $F_{IDEAL} = 0.00785 \text{ m}^2$). For the model, the averaged value of F was further used. Then the resultant ODE for heated water balance reads:

$$q_v(t) = k\sqrt{h(t)} + F \frac{dh(t)}{dt} \quad (6)$$

for some initial (steady-state) water level $h(0) = h^s$. Energy balance of the heated water provides:

$$q_v \rho c_p T_v + \eta P(t) = q \rho c_p T(t) + \alpha F_o [T(t) - T_o(t)] + \alpha_c F_c [T(t) - T_c(t)] + V \rho c_p \frac{dT(t)}{dt} \quad (7)$$

for some initial temperature $T(0) = T^s$, while the simplified energy balance of the cooling water gives:

$$q_c(t) \rho_c c_{pc} T_{cv} + \alpha_c F_c [T(t) - T_c(t)] = q_c(t) \rho_c c_{pc} T_c(t) + V_c \rho_c c_{pc} \frac{dT_c(t)}{dt} \quad (8)$$

for some initial cooling water temperature $T_C(0) = T_C^S$. In the equation above, the volume of cooling water was measured experimentally as $V_C = 30 \text{ ml} = 0.00003 \text{ m}^3$ approx. It is obvious that heat transfer surface areas F_O (heated water \leftrightarrow surroundings) and F_C (cooling water \leftrightarrow heated water) are functions of the water level $h(t)$, i.e. $F_O = \pi D h(t)$ and $F_C = F_C(h)$ generally (this constant is approximated further using experiments). Then the basic set of ODEs describing the heat exchanger in a suitable form for numerical solution can be expressed as:

$$\begin{aligned} \frac{dh(t)}{dt} &= \frac{q_V}{F} - \frac{k\sqrt{h(t)}}{F}, \\ \frac{dT(t)}{dt} &= \frac{q_V T_V}{Fh(t)} + \frac{\eta P(t)}{F\rho c_p h(t)} - \frac{qT(t)}{Fh(t)} - \frac{\alpha\pi D}{F\rho c_p} T(t) \\ &+ \frac{\alpha\pi D T_O}{F\rho c_p} - \frac{\alpha_C F_C(h)T(t)}{F\rho c_p h(t)} + \frac{\alpha_C F_C(h)T_C(t)}{F\rho c_p h(t)}, \quad (9) \\ \frac{dT_C(t)}{dt} &= \frac{T_{CV}}{V_C} q_C(t) + \frac{\alpha_C F_C(h)}{V_C \rho_C c_{PC}} T(t) \\ &- \frac{\alpha_C F_C(h)}{V_C \rho_C c_{PC}} T_C(t) - \frac{1}{V_C} q_C(t) T_C(t), \end{aligned}$$

where the input, state and output variables are denoted as time-dependent. These equations represent a model of dynamics and consequently are described by a set of ordinary differential equations. Steady-state models are also useful as they can be used to assess static properties, e.g. present nonlinearities, suitable working areas, or they can be used to identify unknown parameters experimentally. The same idea was used here to estimate the parameters k , F_C , α and α_C . First, a steady-states model of (9) was obtained by putting all time-derivatives equal to zero and denoting the time-dependent variables as steady using the superscript “s”:

$$\begin{aligned} q_V^S &= q^S = k\sqrt{h^S} \\ \eta P^S &= \alpha F_O (T^S - T_O^S) + \alpha_C F_C(h^S) (T^S - T_C^S) \quad (10) \\ q_C^S \rho_C c_{PC} T_{CV} + \alpha_C F_C(h^S) (T^S - T_C^S) &= q_C^S \rho_C c_{PC} T_C^S \end{aligned}$$

The first equation can be used to identify the valve constant k experimentally. For several steady input flow rates q_V^S the steady-state water level in the vessel h^S was measured and averaged with the following approximate result: $k = 0.128 \text{ m}^{5/2}/\text{s}$. The second and third equations were used to estimate experimentally the heat transfer coefficient α , and the term $\alpha_C F_C$ which is a function of the water level h^S . Several repeated experiments were performed in usual operating points. The results were averaged and processed further to estimate the unknown parameters. The resultant approximation of α was $\alpha = 13.9 \text{ J}\cdot\text{m}^{-2}\cdot\text{K}^{-1}\cdot\text{s}^{-1}$. The other unknown term $\alpha_C F_C$ was obtained as a function of the water level in the form of the following graph, Fig. 8, which has been approximated by a polynomial regression.

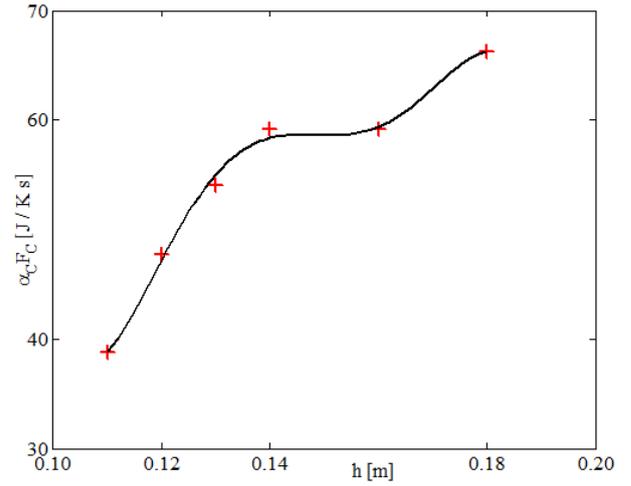


Figure 8: Approximation of the Term $\alpha_C F_C$

The resultant regression equation reads:

$$\begin{aligned} \alpha_C F_C &= -292269698h^5 + 213321855h^4 \\ &- 61608458h^3 + 8792620h^2 - 619260h + 17242. \end{aligned} \quad (11)$$

This equation was further used in the resultant simulation model which was implemented in the MATLAB/Simulink environment and it is formed using the PSV model, described earlier in the paper, together with the heat exchanger model explained in this section.

SIMULATION RESULTS

The developed model was tested in various experimental conditions usual in the course. The tests comprised both open and closed-loop experiments for which suitable controllers were proposed. All the main experiments, linearization, control-oriented analysis and design together with corresponding results are included in (Macek 2015). Here, due to the limited space, only some of them are presented. Comparison of open-loop responses of the model and real-time measurements on the PSV are presented in Fig. 9 and Fig. 10 where the latter one is zoomed in to enable closer inspection.

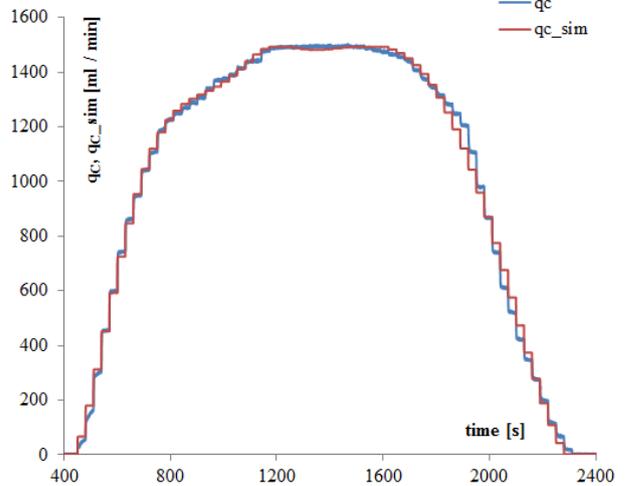


Figure 9: PSV – Open-Loop Comparison of Responses

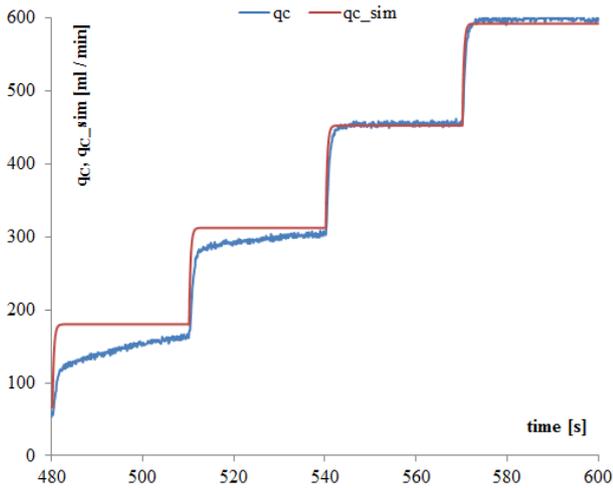


Figure 10: PSV – Detailed Open-Loop Comparison

Overall behaviour during the measurements of the static characteristics presented in Fig. 9 shows relatively good static approximation of the valve. Detailed inspection of the responses in Fig 10 reveals lower accuracy in the dynamic part of the model, especially in the lower flow-rates, which is given by the only one (averaged) simple approximation of the dynamics (3) in the whole working range of the valve. Closed-loop responses obtained using an auxiliary PI-controller are presented in Fig. 11.

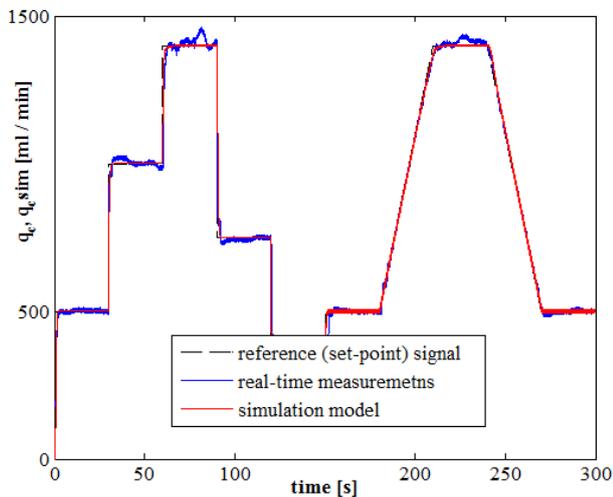


Figure 11: PSV – Closed-Loop Comparison

Here, it is clear that the model approximates the real valve in the closed loop system well and can be further used for simulation experiments. The heat exchanger model presented in the previous section has been also tested properly in both open and closed-loop settings. Several open-loop responses are presented further, other (especially closed-loop ones) are beyond the space limit of this contribution and can be found in (Macek 2015). Next graphs present open-loop dynamic responses for the following usual experimental conditions: $h = 20 \text{ cm}$, $q_C = 1200 \text{ ml/min}$ and $P = 5 \%$, in different regimes. The first one (Fig. 12) presents behaviour without any heating and cooling, i.e. it is a process of heating the

water inside the vessel to ambient temperature T_O . From the graph it is clear how the process can be slow – with responses in hours. The second graph, Fig. 13, present the case of heating with prescribed power ($P = 5 \%$).

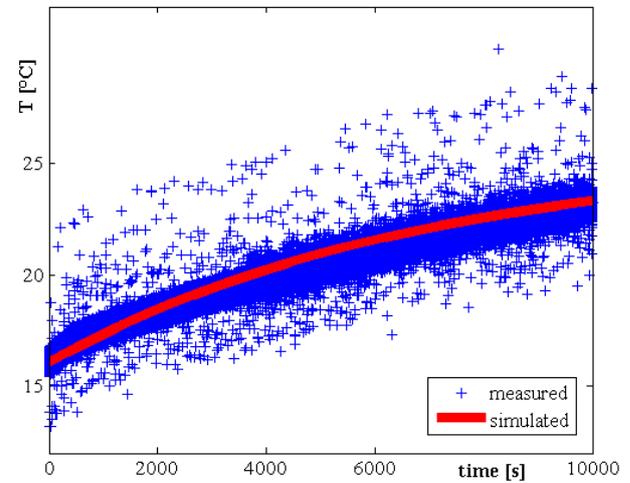


Figure 12: Heat Exchanger – Open-Loop Comparison of Responses (Heating & Cooling OFF)

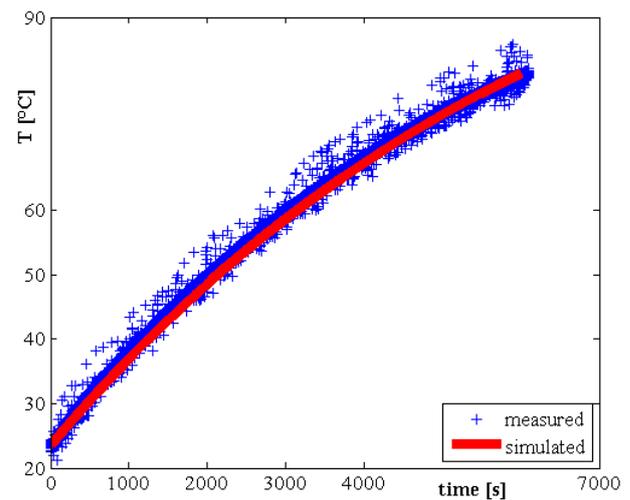


Figure 13: Heat Exchanger – Open-Loop Comparison of Responses (Heating ON)

Both responses reveal good model tracking of the real-time data. Next responses present behaviour in the cooling case – Fig. 14 for cooling only and Fig. 15 for simultaneous heating and cooling. The cooling comparison (Fig. 14) reveals a small acceptable mismatch in the simulation model for cooling. The last graph, Fig. 15, shows high noise ratio during the measurements which has led to the need of data filtration which has been implemented simply using a first order low-pass filter.

The overall results indicate that the developed simulation model, although simple, can be successfully used for simulation experiments comparable to real-time measurements when there is no need for high accuracy in terms of absolute values, which is the common case of control-oriented simulation.

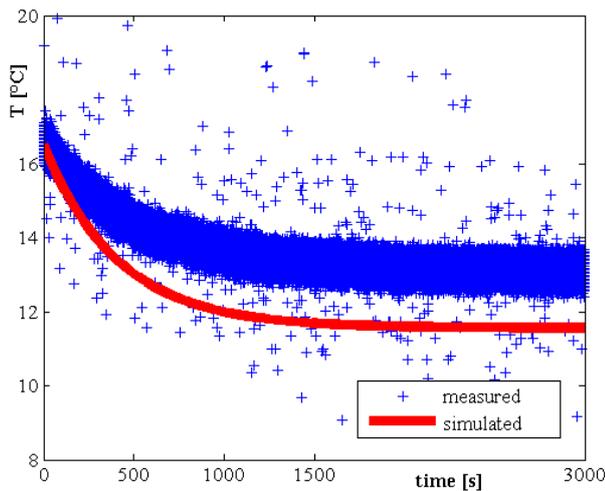


Figure 14: Heat Exchanger – Open-Loop Comparison of Responses (Cooling ON)

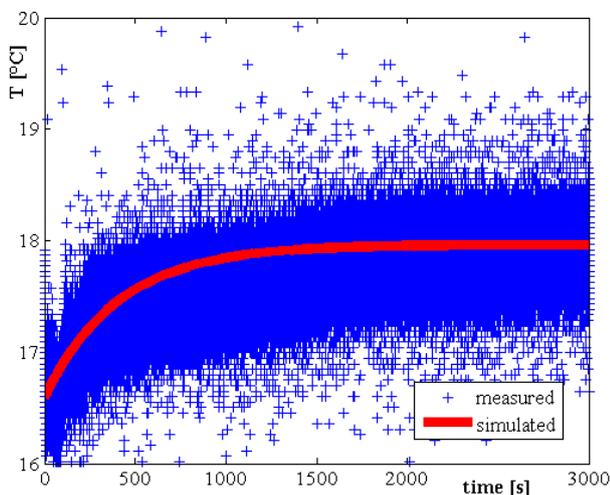


Figure 15: Heat Exchanger – Open-Loop Comparison of Responses (Heating & Cooling ON)

CONCLUSIONS

This paper has presented a simulation model of the PCT40 heat exchanger. It was developed using both analytical and experimental methods – the basic model structure was obtained analytically and its unknown parameters were identified using repeated real-time measurements and corresponding data processing. The resultant model, though relatively simple, exhibits good fit in terms of main dynamic properties, which is sufficient for the outlined purposes, i.e. control-oriented simulation. As a result, both students and teachers working in the laboratory can save time during their experiments – they can easily and quickly test their control algorithms prior to the real-time implementation. The whole system could be, of course, modelled more precisely, especially the cooling part, resulting in a distributed-parameters system described by partial differential equations (PDEs). However, this was not the goal of the presented modelling & simulation study and

can be a theme for further work, comparing the presented simplified model with the more precise, distributed one.

REFERENCES

- Amira. 2016. *Laboratory Equipment for Research and Practicals of Control Engineering* [online]. Available at: http://www.ict.com.tw/AI/Amira/amira/home_e.htm
- Armfield. 2005. *PCT40 Multifunction Process Control Teaching System: Basic Process Control Unit*. Instruction Manual. Armfield Limited, Ringwood, England.
- Armfield. 2016. *Armfield Engineering Teaching Equipment - Education, teaching and research equipment for chemical engineering, mechanical engineering, civil engineering and all the major disciplines* [online]. Available at: <http://discoverarmfield.com>
- Chalupa, P.; J. Novak and V. Bobal. 2012. “Comprehensive Model of DTS200 Three Tank System in Simulink.” *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol.6, No.2, 358-365.
- Faculty of Applied Informatics, Tomas Bata University in Zlín. 2016. [online]. Available at: <http://www.utb.cz/fai-en>
- Feedback. 2016. *Feedback* [online]. Available at: <http://www.feedback-instruments.com>
- LD Didactic. 2016. *Leybold – LD Didactic* [online]. Available at: <http://www.ld-didactic.de/en>
- Ljung L. and G. Torkel. 1994. *Modeling of Dynamic Systems*. Prentice Hall, New Jersey.
- Macek, D. 2015. *Modelling a PCT40 Heat Exchanger*. Master’s thesis. Tomas Bata University in Zlín, CZ.
- Severance, F.L. 2001. *System Modeling and Simulation: An Introduction*. Wiley, Chichester.
- TecEquipment. 2016. *Control Engineering Teaching Equipment from TecEquipment Ltd* [online]. Available at: <http://www.tecequipment.com/Control.aspx>
- Wellstead, P.E. 1979. *Introduction to Physical System Modelling*. Academic Press, London.

AUTHOR BIOGRAPHIES



FRANTIŠEK GAZDOŠ was born in Zlín, Czech Republic in 1976, and graduated from the Brno University of Technology in 1999 with MSc. degree in Automation. He then followed studies of Technical Cybernetics at Tomas Bata

University in Zlín, obtaining Ph.D. degree in 2004. He became Associate Professor for Machine and Process Control in 2012 and now works in the Department of Process Control, Faculty of Applied Informatics of Tomas Bata University in Zlín, Czech Republic.

He is author or co-author of more than 80 journal contributions and conference papers giving lectures at foreign universities, such as Politecnico di Milano, University of Strathclyde Glasgow, Universidade Técnica de Lisboa and others. His research cover the area of process modelling, simulation and control. His e-mail address is: gazdos@fai.utb.cz.

PREDICTIVE CONTROL OF THREE-TANK-SYSTEM UTILIZING BOTH STATE-SPACE AND INPUT-OUTPUT MODELS

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

Predictive control, Adaptive control, Three-Tank-System, Recursive identification, State-Space model.

ABSTRACT

The paper introduces a controller which integrates a predictive control synthesis based on a multivariable state-space model of a controlled system and an on-line identification of an ARX model corresponding to the state-space model. The used approach then combines both the state-space and input-output models. The model parameters are recursively estimated using the recursive least squares method. The control algorithm is based on the Generalised Predictive Control (GPC) method. The optimization was realized by minimization of a quadratic objective function. The controller was applied for simulation control of a model of a three-tank-system. The objective simulation model is a two input-two output (TITO) nonlinear system. Results of simulations are also included.

INTRODUCTION

Typical technological processes require the simultaneous control of several variables related to one system. Each input may influence all system outputs. The three – tank – system in Fig. 1 is a typical multivariable nonlinear system with significant cross – coupling. The design of a controller for such a system must be quite sophisticated if the system is to be controlled adequately. Simple decentralized PI or PID controllers largely do not yield satisfactory results. There are many different advanced methods of controlling multi-input–multi-output (MIMO) systems. The problem of selecting an appropriate control technique often arises. Perhaps the most popular way of controlling MIMO processes is by designing decoupling compensators to suppress the interactions (e.g. Krishnawamy et al, 1991) and the designing multiple SISO controllers (e.g. Luyben, 1986). This requires determining how to pair the controlled and manipulated variables. One of the most effective approaches to control of multivariable systems is model predictive control (MPC) (Camacho and Bordons, 2004, Morari and Lee, 1999, Bitmead et al, 1990). An advantage of model predictive control is that multivariable systems can be handled in a straightforward manner. When using most of other approaches, the control actions are

taken based on past errors. MPC uses also future values of the reference signals.

The aim of this contribution is implementation of an adaptive predictive controller for control of the simulation three – tank – system model. The design of the controller is based on a state – space model. An initial state – space model was constructed according to first principles and physical rules. The three-tank-system is a nonlinear system with variable parameters and its description by a linear model is valid only in a neighbourhood of a steady state. Self-tuning controllers (Bobál et al, 2005) are a possible approach to the control of this kind of system. However, the state – space description is not quite suitable for a recursive identification of the parameters of the process which is performed during control with self – tuning controllers. The state space model was then converted to a model in the form of difference equations. This model is suitable for the recursive identification. So the proposed approach combines both types of models. The state – space model is used for the controllers design and the corresponding input/output model for the estimation of the unknown parameters. Of course it is possible to base the controllers design on the input/output model as well. But using of state-space model enables to solve tasks which are unsolvable when using an input/output model. For example control with state constraints.

Reverse conversion of the difference equations to the original state – space model is not possible. An alternative state – space model was then established and used for the controllers design. This model corresponds to the original model despite the fact that it has a different structure. So it is possible to assume that this model describes main properties of the controlled process as well as the original model.

The Generalised Predictive Control (GPC) method (Clarke et al, 1987 I., Clarke et al, 1987 II.) was then applied for the controllers design. In the optimization part of the algorithm a quadratic cost function was used. The algorithm takes into account constraints of manipulated variables. The recursive least squares method is used in the identification part.

MODEL OF INTERCONNECTED TANKS

The three – tank – system can be viewed as a prototype of many industrial applications in process industry, such as chemical and petrochemical plants, oil and gas systems. It is desirable to obtain the tank levels at

certain values. The principle scheme of the model is shown in Fig 1. It consists of three tanks numbered from left to right as T1, T2 and T3. These are connected serially with each other by cylindrical pipes. Liquid, which is collected in a reservoir, is pumped into the first and the third tanks to maintain their levels. The level in the tank T2 is a response which is uncontrollable. It affects the level in the two end tanks.

Q_1 and Q_2 are the flow rates of the pumps 1 and 2. The tanks are connected serially by valves V1 and V2. Valves V3 and V4 represent leakage from tanks T1 and T3.

The model was controlled as a two input – two output (TITO) system. The outputs are controllable liquid levels of the tanks T1 and T2 and the inputs are the pump flow rates Q_1 and Q_2 . Each pump flow rate affects both liquid levels. This is the coupling. The systems inputs and outputs interact and the whole system is a multivariable system.

The three – tank – system is a nonlinear system with variable parameters. Flow rate of liquid through a valve is proportional to square root of a pressure difference in front of and behind the valve.

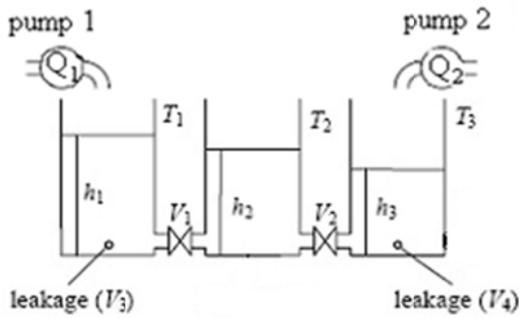


Figure 1: Principle scheme of three-tank-system

If we consider flow rates balances we get dynamic mathematical model as a set of three ordinary differential equations which can be written in form (1). We can define following parameters: section of cylinder S (the tanks are supposed to have equal sections), and liquid levels in particular tanks h_1, h_2, h_3 . We will consider two input variables which are flow rates of the pumps Q_1 and Q_2 , two output variables represented by liquid levels of the two outer tanks h_1 and h_3 and three state variables which are the liquid levels h_1, h_2, h_3 .

$$\begin{aligned} S \frac{dh_1(t)}{dt} &= Q_1(t) - \text{sign}[h_1(t) - h_2(t)]q_1(t) - q_3(t) \\ S \frac{dh_2(t)}{dt} &= \text{sign}[h_1(t) - h_2(t)]q_1(t) - \text{sign}[h_2(t) - h_3(t)]q_2(t) \\ S \frac{dh_3(t)}{dt} &= Q_2(t) + \text{sign}[h_2(t) - h_3(t)]q_2(t) - q_4(t) \end{aligned} \quad (1)$$

Flow rate of liquid through a valve is proportional to square root of a pressure difference in front of and behind the valve. Particular flow rates in our case are then given by following equations.

$$\begin{aligned} q_1(t) &= k_1 \sqrt{|h_1(t) - h_2(t)|} & q_2(t) &= k_2 \sqrt{|h_2(t) - h_3(t)|} \\ q_3(t) &= k_3 \sqrt{h_1(t)} & q_4(t) &= k_4 \sqrt{h_3(t)} \end{aligned} \quad (2)$$

Where k_1, k_2, k_3 and k_4 are coefficients of the valves. The model then takes following form of nonlinear equations which express relations among state variables.

$$\begin{aligned} S \frac{dh_1(t)}{dt} &= Q_1(t) - \text{sign}[h_1(t) - h_2(t)]k_1 \sqrt{|h_1(t) - h_2(t)|} - k_3 \sqrt{h_1(t)} \\ S \frac{dh_2(t)}{dt} &= \text{sign}[h_1(t) - h_2(t)]k_1 \sqrt{|h_1(t) - h_2(t)|} - \text{sign}[h_2(t) - h_3(t)]k_2 \sqrt{|h_2(t) - h_3(t)|} \\ S \frac{dh_3(t)}{dt} &= Q_2(t) + \text{sign}[h_2(t) - h_3(t)]k_2 \sqrt{|h_2(t) - h_3(t)|} - k_4 \sqrt{h_3(t)} \end{aligned} \quad (3)$$

On the basis of equations (3) was realized a model in the Matlab/Simulink environment in the form of S-function. Measurement of the static characteristics was simulated. The system parameters used in the simulation are in Table 1.

Table1. Three-Tank-System parameters

Tank cross section area S	$S = 7 \text{ m}^2$
Valve coefficient k_1	$k_1 = 5 \text{ m}^{5/2}/\text{min}$
Valve coefficient k_2	$k_2 = 4 \text{ m}^{5/2}/\text{min}$
Valve coefficient k_3	$k_3 = 3 \text{ m}^{5/2}/\text{min}$
Valve coefficient k_4	$k_4 = 2 \text{ m}^{5/2}/\text{min}$

The static characteristics are shown in Figure 2.

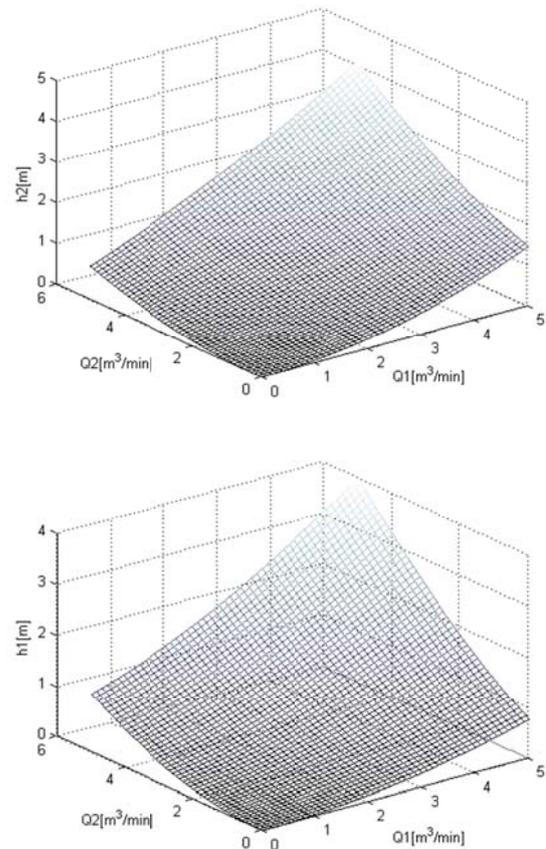


Figure 2: Static characteristics of three-tank-system

Initial conditions in equations (1) we can obtain by solving of a steady state model. In the steady state holds

$$\begin{aligned} Q_1^s &= q_1^s - q_3^s & q_1^s &= q_2^s \\ Q_2^s &= q_2^s - q_4^s \end{aligned} \quad (4)$$

After substitution of (4) to (3) we can obtain expressions for computation of steady state liquid levels and then position of the operational point (h_1^s, h_2^s, h_3^s)

$$\begin{aligned} q_1^s &= k_1 \sqrt{h_1^s - h_2^s} & q_2^s &= k_2 \sqrt{h_2^s - h_3^s} \\ q_3^s &= k_3 \sqrt{h_1^s} & q_4^s &= k_4 \sqrt{h_3^s} \end{aligned} \quad (5)$$

In a steady state always holds $h_1^s > h_2^s$ $h_2^s > h_3^s$

We can compute a linearized mathematical model which is a differential model. Let us establish differences of liquid levels and input flow rates from the initial steady state as

$$x_j(t) = \Delta h_j(t) = h_j(t) - h_j^s \quad u_j(t) = \Delta Q_j(t) = Q_j(t) - Q_j^s \quad (6)$$

Now we transcribe equations (1) to the differential form

$$\begin{aligned} S \frac{d\Delta h_1(t)}{dt} &= \Delta Q_1(t) - \text{sign}[h_1(t) - h_2(t)] \Delta q_1(t) - \Delta q_3(t) \\ S \frac{d\Delta h_2(t)}{dt} &= \text{sign}[h_1(t) - h_2(t)] \Delta q_1(t) - \text{sign}[h_2(t) - h_3(t)] \Delta q_2(t) \\ S \frac{d\Delta h_3(t)}{dt} &= \Delta Q_2(t) + \text{sign}[h_2(t) - h_3(t)] \Delta q_2(t) - \Delta q_4(t) \end{aligned} \quad (7)$$

The differences are then substituted by linear terms of their Taylor polynomial in the neighbourhood of the operational point (h_1^s, h_2^s, h_3^s)

$$\Delta q_1(t) \approx \left(\frac{\partial q_1(t)}{\partial h_1(t)} \right)^s \Delta h_1(t) + \left(\frac{\partial q_1(t)}{\partial h_2(t)} \right)^s \Delta h_2(t) = \quad (8)$$

$$= \frac{q_1^s}{2(h_1^s - h_2^s)} (\Delta h_1(t) - \Delta h_2(t)) = K_1 (\Delta h_1(t) - \Delta h_2(t))$$

$$\Delta q_2(t) \approx \left(\frac{\partial q_2(t)}{\partial h_2(t)} \right)^s \Delta h_2(t) + \left(\frac{\partial q_2(t)}{\partial h_3(t)} \right)^s \Delta h_3(t) = \quad (9)$$

$$= \frac{q_2^s}{2(h_2^s - h_3^s)} (\Delta h_2(t) - \Delta h_3(t)) = K_2 (\Delta h_2(t) - \Delta h_3(t))$$

$$\Delta q_3(t) \approx \left(\frac{dq_3(t)}{dh_1(t)} \right)^s \Delta h_1(t) = \frac{q_3^s}{2h_1^s} \Delta h_1(t) = K_3 \Delta h_1(t) \quad (10)$$

$$\Delta q_4(t) \approx \left(\frac{dq_4(t)}{dh_3(t)} \right)^s \Delta h_3(t) = \frac{q_4^s}{2h_3^s} \Delta h_3(t) = K_4 \Delta h_3(t) \quad (11)$$

The coefficients K_1, K_2, K_3, K_4 are dependent on the operational point position. After substitution of (8), (9), (10) and (11) to (7) we obtain the linearized differential model of the system in the form

$$\begin{aligned} S \frac{dx_1(t)}{dt} &= u_1(t) - K_1(x_1(t) - x_2(t)) - K_3 x_1(t) \\ S \frac{dx_2(t)}{dt} &= K_1(x_1(t) - x_2(t)) - K_2(x_2(t) - x_3(t)) \\ S \frac{dx_3(t)}{dt} &= u_2(t) + K_2(x_2(t) - x_3(t)) - K_4 x_3(t) \end{aligned} \quad (12)$$

with zero initial conditions. The model can be transcribed to

$$\begin{aligned} \frac{dx_1(t)}{dt} &= a_{11}x_1(t) + a_{12}x_2(t) + b_{11}u_1(t) \\ \frac{dx_2(t)}{dt} &= a_{21}x_1(t) + a_{22}x_2(t) + a_{23}x_3(t) \\ \frac{dx_3(t)}{dt} &= a_{32}x_2(t) + a_{33}x_3(t) + b_{33}u_2(t) \end{aligned} \quad (13)$$

where

$$\begin{aligned} a_{11} &= \frac{-K_1 - K_3}{S} & a_{12} &= \frac{K_1}{S} & b_{11} &= \frac{1}{S} \\ a_{21} &= \frac{K_1}{S} & a_{22} &= \frac{-K_1 - K_2}{S} & a_{23} &= -\frac{K_1}{S} \\ a_{32} &= \frac{K_2}{S} & a_{33} &= \frac{-K_2 - K_4}{S} & b_{33} &= \frac{1}{S} \end{aligned} \quad (14)$$

The state equations can be transcribed to a matrix form

$$\begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \\ \frac{dx_3(t)}{dt} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} b_{11} & 0 \\ 0 & 0 \\ 0 & b_{33} \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \quad (15)$$

The output equation can be defined as

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_1(t) \\ h_2(t) \\ h_3(t) \end{bmatrix} \quad (16)$$

The continuous – time process model can be transferred for a given sampling time T_v to a discrete time state – space model

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) \end{aligned} \quad (17)$$

The discrete state – space model which structure corresponds to the continuous – time state space model (15) and (16) takes the following general form

$$\begin{aligned} \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix} &= \begin{bmatrix} A_1 & A_2 & A_3 \\ A_4 & A_5 & A_6 \\ A_7 & A_8 & A_9 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix} + \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \\ B_5 & B_6 \end{bmatrix} \begin{bmatrix} u_1(k) \\ u_2(k) \end{bmatrix} \\ \begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix} \end{aligned} \quad (18)$$

The model has 15 unknown parameters. This state – space model, which is in fact based on first principles and perceives physical nature of the process, can be transcribed to difference equations

$$\begin{aligned} y_1(k) &= a_1 y_1(k-1) + a_2 y_1(k-2) + a_3 y_2(k-1) + \\ &+ a_4 y_2(k-2) + b_1 u_1(k-1) + b_2 u_1(k-2) + \\ &+ b_3 u_2(k-1) + b_4 u_2(k-2) \\ y_2(k) &= a_5 y_1(k-1) + a_6 y_1(k-2) + a_7 y_2(k-1) + \\ &+ a_8 y_2(k-2) + b_5 u_1(k-1) + b_6 u_1(k-2) + \\ &+ b_7 u_2(k-1) + b_8 u_2(k-2) \end{aligned} \quad (19)$$

where

$$\begin{aligned} a_1 &= A_1; a_2 = A_2 A_4 - \frac{A_2 A_5 A_7}{A_8}; a_3 = \frac{A_2 A_5}{A_8} + A_3; \\ a_4 &= A_2 A_6 - \frac{A_2 A_5 A_9}{A_8}; b_1 = B_1; b_2 = A_2 B_3 - \frac{A_2 A_5 B_5}{A_8} \\ b_3 &= B_2; b_4 = A_2 B_4 - \frac{A_2 A_5 B_6}{A_8} \\ a_5 &= \frac{A_8 A_5}{A_2} + A_7; a_6 = A_8 A_4 - \frac{A_8 A_5 A_1}{A_2}; a_7 = A_9; \\ a_8 &= A_8 A_6 - \frac{A_8 A_5 A_3}{A_2}; b_5 = B_5; b_6 = A_8 B_3 - \frac{A_8 A_5 B_1}{A_2} \\ b_7 &= B_6; b_8 = A_8 B_4 - \frac{A_8 A_5 B_2}{A_2} \end{aligned} \quad (20)$$

This model is suitable for a recursive identification of the unknown parameters of the process and was used in the identification part. From the equations (7) it is obvious that conversion of the obtained difference equations to the original state – space form is not possible. The difference equations were then converted to an alternative state – space model. This model corresponds to the original model despite the fact that it has a different structure. So it is possible to assume that this model describes main nature of the controlled process as well as the original model.

New state variables were established. The model has four state variables defined as follows

$$\begin{aligned} x_1(k) &= y_1(k) \\ x_2(k) &= y_1(k+1) - b_1 u_1(k) \\ x_3(k) &= y_2(k) \\ x_4(k) &= y_2(k+1) - b_3 u_2(k) \end{aligned} \quad (21)$$

The state – space model then takes the form

$$\begin{aligned} \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ a_2 & a_1 & a_4 & a_3 \\ 0 & 0 & 0 & 1 \\ a_6 & a_5 & a_8 & a_7 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix} + \\ &+ \begin{bmatrix} b_1 & b_3 \\ a_1 b_1 + a_3 b_5 + b_2 & a_1 b_3 + a_3 b_7 + b_4 \\ b_5 & b_7 \\ a_5 b_1 + a_7 b_5 + b_6 & a_5 b_3 + a_7 b_7 + b_8 \end{bmatrix} \begin{bmatrix} u_1(k) \\ u_2(k) \end{bmatrix} \\ \begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{bmatrix} \end{aligned} \quad (22)$$

For purposes of the controller design it was necessary to incorporate an integrator to the model of the process in order to achieve zero permanent control error. One possibility is to define a new state vector by making $u(k-1)$ an additional internal state.

$$\bar{x}(k) = \begin{pmatrix} x(k) \\ u(k-1) \end{pmatrix} \quad (23)$$

Then we can obtain an augmented state space – model in the form

$$\begin{aligned} \bar{x}(k+1) &= \begin{pmatrix} A & B \\ 0 & I \end{pmatrix} \bar{x}(k) + \begin{pmatrix} B \\ I \end{pmatrix} \Delta u(k) = \bar{A} \bar{x}(k) + \bar{B} \Delta u(k) \\ y(k) &= (C \ 0) \bar{x}(k) = \bar{C} \bar{x}(k) \end{aligned} \quad (24)$$

$$\begin{aligned} \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \\ u_1(k) \\ u_2(k) \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 & b_1 & b_3 \\ a_2 & a_1 & a_4 & a_3 & a_1 b_1 + a_3 b_5 + b_2 & a_1 b_3 + a_3 b_7 + b_4 \\ 0 & 0 & 0 & 1 & b_5 & b_7 \\ a_6 & a_5 & a_8 & a_7 & a_5 b_1 + a_7 b_5 + b_6 & a_5 b_3 + a_7 b_7 + b_8 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ u_1(k-1) \\ u_2(k-1) \end{bmatrix} + \\ &+ \begin{bmatrix} b_1 & b_3 \\ a_1 b_1 + a_3 b_5 + b_2 & a_1 b_3 + a_3 b_7 + b_4 \\ b_5 & b_7 \\ a_5 b_1 + a_7 b_5 + b_6 & a_5 b_3 + a_7 b_7 + b_8 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta u_1(k) \\ \Delta u_2(k) \end{bmatrix} \\ \begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ u_1(k-1) \\ u_2(k-1) \end{bmatrix} \end{aligned} \quad (25)$$

This model was then used for the controller's design.

DESIGN OF PREDICTIVE CONTROLLER

The basic idea of MPC is to use a model of a controlled process to predict N future outputs of the process. A trajectory of future manipulated variables is given by solving an optimization problem incorporating a suitable cost function and constraints. Only the first element of the obtained control sequence is applied. The whole procedure is repeated in following sampling period. This principle is known as the receding horizon strategy. The computation of a control law of MPC is based on minimization of the following criterion

$$J(k) = \sum_{j=N_u}^N e(k+j)^2 + \lambda \sum_{j=1}^{N_u} \Delta u(k+j)^2 \quad (26)$$

where $e(k+j)$ is a vector of predicted control errors, $\Delta u(k+j)$ is a vector of future increments of manipulated variables (for the system with two inputs and two outputs each vector has two elements), N is length of the prediction horizon, N_u is length of the control horizon and λ is a weighting factor of control increments.

A predictor in a vector form is given by

$$\hat{y} = \mathbf{G}\Delta u + y_0 \quad (27)$$

Where \hat{y} is a vector of system predictions along the horizon of the length N , Δu is a vector of control increments over the horizon N_u , y_0 is the free response vector. \mathbf{G} is a matrix of the dynamics given as

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_0 & 0 & \cdots & \cdots & 0 \\ \mathbf{G}_1 & \mathbf{G}_0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \mathbf{G}_0 & 0 \\ \mathbf{G}_{N-1} & \cdots & \cdots & \cdots & \mathbf{G}_0 \end{bmatrix} \quad (28)$$

where sub-matrices \mathbf{G}_i have dimension 2x2 and contain values of the step sequence.

Predictions over the horizon N are computed recursively using (14), resulting in

$$\hat{y}(k+j) = \overline{\mathbf{CA}}^j \hat{x}(k) + \sum_{i=0}^{j-1} \overline{\mathbf{CA}}^{j-i-1} \overline{\mathbf{B}} \Delta u(k+i) \quad (29)$$

where $\hat{x}(k)$ is an estimation of the state vector $x(k)$.

The equation (27) can be written after substitution as

$$\hat{y} = \mathbf{F}\hat{x}(k) + \mathbf{G}\Delta u \quad (30)$$

The free and forced responses are then computed recursively

$$\mathbf{F}_{ij} = \overline{\mathbf{CA}}^j \quad \mathbf{F} = \begin{bmatrix} \overline{\mathbf{CA}} \\ \overline{\mathbf{CA}}^2 \\ \vdots \\ \overline{\mathbf{CA}}^{N_s} \end{bmatrix} \quad \mathbf{G}_{ij} = \overline{\mathbf{CA}}^{i-j} \overline{\mathbf{B}}$$

$$\mathbf{G} = \begin{bmatrix} \overline{\mathbf{CB}} & 0 & 0 & \cdots \\ \overline{\mathbf{CAB}} & \overline{\mathbf{CB}} & 0 & \cdots \\ \overline{\mathbf{CA}^2\mathbf{B}} & \overline{\mathbf{CAB}} & \overline{\mathbf{CB}} & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ \overline{\mathbf{CA}^{N_s-1}\mathbf{B}} & \overline{\mathbf{CA}^{N_s-2}\mathbf{B}} & \overline{\mathbf{CA}^{N_s-3}\mathbf{B}} & \cdots \end{bmatrix} \quad (31)$$

The criterion (26) can be written in a general vector form

$$J = (\hat{y} - \mathbf{w})^T (\hat{y} - \mathbf{w}) + \lambda \Delta u^T \Delta u \quad (32)$$

where \mathbf{w} is a vector of the reference trajectory. The criterion can be modified using the expression (30) to

$$J = 2\mathbf{g}^T \Delta u + \Delta u^T \mathbf{H} \Delta u \quad (33)$$

where the gradient \mathbf{g} and the Hess matrix \mathbf{H} are defined by following expressions

$$\mathbf{g}^T = \mathbf{G}^T (\mathbf{F}\hat{x}(k) - \mathbf{w}) \quad \mathbf{H} = \mathbf{G}^T \mathbf{G} + \lambda \mathbf{I} \quad (34)$$

In case of the three – tank – system, manipulated variables have a limited range of action. MPC can consider constrained input and output signals in the process of the controller design (Maciejowski, 2002). This is one of the major advantages of predictive control. General formulation of predictive control with constraints is then as follows

$$\min_{\Delta u} 2\mathbf{g}^T \Delta u + \Delta u^T \mathbf{H} \Delta u \quad (35)$$

owing to

$$\mathbf{A}\Delta u \leq \mathbf{b} \quad (36)$$

The inequality (36) expresses the constraints in a compact form.

The optimization problem is then solved numerically by quadratic programming in each sampling period. The first element of the resulting vector is then applied as the increment of the manipulated variable.

The design of the controller is based on the state – space model which has significant advantages in predictive control. However, this model is not suitable for recursive estimation of its parameters. This requirement suits an input – output model. A problem is that it is not possible to simply convert a state – space model to an input – output model and vice versa. An alternative possible conversion which enables both controllers design based on the state – space model and simple recursive estimation of its parameters is presented in the previous section.

RECURSIVE IDENTIFICATION

The control algorithm was applied as a self-tuning controller (as discussed in section 1). The unknown parameters of the controlled process were identified on the basis of the model in the form of difference equations (19) which are suitable for recursive identification. Self-tuning control is based on the online identification of a model of a controlled process. Each self – tuning controller consists of an on – line identification part and a control part.

Various discrete linear models are used to describe dynamic behaviour of controlled systems; see for example the overview in (Nelles, 2001). The most widely applied linear dynamic model is the ARX model. Usually the ARX model is tested first and more complex model structures are only examined if it does not perform satisfactorily. However, the ARX model matches the structure of many real processes. The parameters can be easily estimated by a linear least-squares technique. It is suitable also for the proposed difference equations (19).

The ARX model describing the TITO process is defined as

$$\begin{aligned} y_1(k) &= \boldsymbol{\theta}_1(k)\boldsymbol{\phi}(k-1) + e_{s_1}(k) \\ y_2(k) &= \boldsymbol{\theta}_2(k)\boldsymbol{\phi}(k-1) + e_{s_2}(k) \end{aligned} \quad (37)$$

where $e_{s_1}(k)$, $e_{s_2}(k)$ are non-measurable disturbances. Parameter vectors are specified as follows:

$$\begin{aligned} \boldsymbol{\theta}_1^T(k) &= [a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4] \\ \boldsymbol{\theta}_2^T(k) &= [a_5, a_6, a_7, a_8, b_5, b_6, b_7, b_8] \end{aligned} \quad (38)$$

The data vector is

$$\boldsymbol{\phi}^T(k-1) = [y_1(k-1), y_1(k-2), y_2(k-1), y_2(k-2), u_1(k-1), u_1(k-2), u_2(k-1), u_2(k-2)] \quad (39)$$

The recursive least squares method (Bobál et al, 2005) was then used for the estimation of the parameters.

SIMULATION EXAMPLE

The model in the form of S-function which was previously introduced was used as a simulated controlled system. Its parameters are in Table 1.

Simulation parameters are in Table 2. The tuning parameters that are the prediction and control horizons, the weighting coefficient λ and sample time were tuned experimentally. There is a lack of clear theory relating to the closed loop behavior to design parameters.

In Figures 3 and 4 are shown time responses of the control when the initial parameter estimates were chosen without any a priori information. The reference signal was chosen as a step function. The controlled and manipulated variables were stabilized and asymptotic tracking of the reference signal was achieved. In Figure 5 are shown courses of the identified parameters of the controlled system. It is obvious that step changes of the reference signal cause changes of the parameters in particular operating points of the nonlinear system.

Table 2. Simulation parameters

Prediction horizon N	$N = 5$
Control horizon N_u	$N_u = 5$
Sample time T_s	0,4 min
Weighting factor λ	0,1
Minimum flow rate pump 1	$Q_{1\min} = 0 \text{ m}^3/\text{min}$
Maximum flow rate pump 1	$Q_{1\max} = 10 \text{ m}^3/\text{min}$
Minimum flow rate pump 2	$Q_{2\min} = 0 \text{ m}^3/\text{min}$
Maximum flow rate pump 2	$Q_{2\max} = 10 \text{ m}^3/\text{min}$

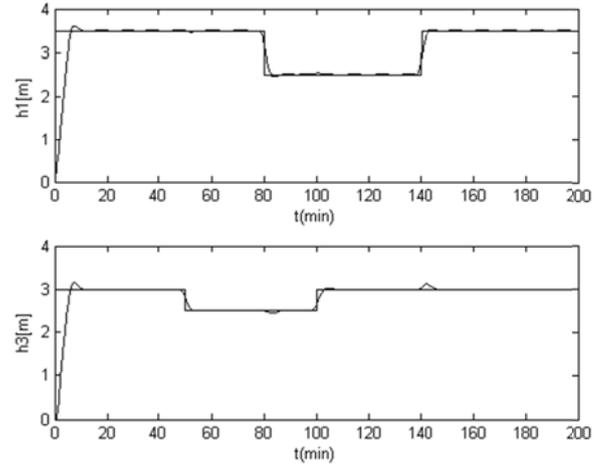


Figure 3: Simulation control of three-tank-system – controlled and reference variables

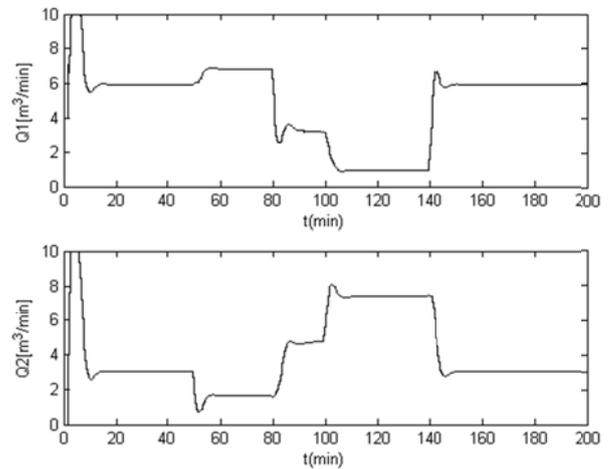


Figure 4: Simulation control of three-tank-system – manipulated variables

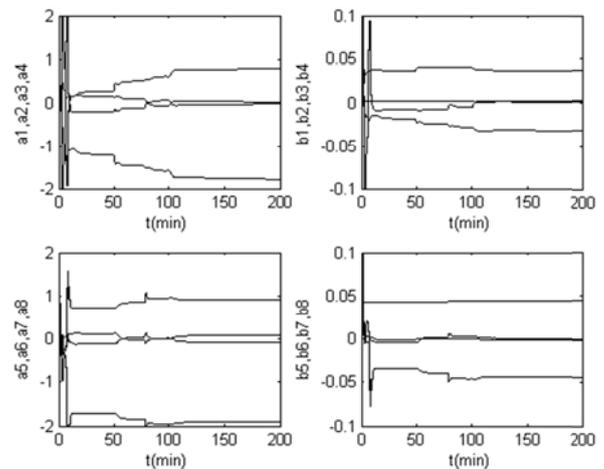


Figure 5: Courses of the estimated parameters

CONCLUSIONS

The model predictive self - tuning controller was proposed and verified by simulation control of the nonlinear time varying system. The proposed approach combines both state – space and input output models of the controlled system. State space models do not enable simple recursive parameters estimation. On the other hand predictive controllers based on state – space formulation are better for handling of multivariable systems and they also enable to solve tasks which are unsolvable when using input/output models. State – space model is then used for the controllers design and the corresponding input/output model for the estimation of the unknown parameters of the process. The original state – space model based on first principles and physical rules was converted to the difference equations. Reverse conversion of the difference equations to the original state – space form was not possible. An alternative state – space model was then established and used for the controllers design. It is possible to assume that this model describes main properties of the controlled process as well as the original model. General principles were elaborated on a specific system with two inputs and two outputs that is often applicable in industrial practice. Control algorithm based on the specific model was derived.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and also by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089 and also by the Programme EEA and Norway Grants for funding via grant on Institutional cooperation project nr. NF-CZ07-ICP-4-345-2016

REFERENCES

- Bitmead, R. R., Gevers, M., Hertz, V. 1990. *Adaptive Optimal Control. The Thinking Man's GPC*, Prentice Hall, Englewood Cliffs, New Jersey.
- Bobal, V., Böhm, J., Fessler, J., Machacek, J. 2005. *Digital Self-Tuning Controllers*, Springer - Verlag, London.
- Camacho, E. F., Bordons, C. 2004. *Model Predictive Control*, Springer-Verlag, London.
- Clarke, D. W., Mohtadi, C., Tuffs, P. S. 1987. Generalized predictive control, part I: the basic algorithm. *Automatica*, 23, 137-148.
- Clarke, D. W., Mohtadi, C., Tuffs, P. S. 1987. Generalized predictive control, part II: extensions and interpretations. *Automatica*, 23, 149-160.
- Krishnawamy, P.R., et al. 1991. Reference System Decoupling for Multivariable Control. *Ind. Eng. Chem. Res.*, 30, 662-670.
- Luyben, W.L. 1986. Simple Method for Tuning SISO Controllers in Multivariable Systems. *Ing. Eng. Chem. Process Des. Dev.*, 25, 654 – 660.
- Maciejowski, J.M. 2002. *Predictive Control with Constraints*, Prentice Hall, London.
- Morari, M., Lee, J. H. 1999. Model predictive control: past, present and future. *Computers and Chemical Engineering*, 23, 667-682.
- Nelles, O. 2001. *Nonlinear System Identification*, Springer-Verlag, Berlin.

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 bbobal@fai.utb.cz.

PREDICTIVE CONTROL OF DIFFERENTIAL DRIVE MOBILE ROBOT CONSIDERING DYNAMICS AND KINEMATICS

Rahul Sharma K.

Daniel Honc

František Dušek

Department of Process control

Faculty of Electrical Engineering and Informatics, University of Pardubice, Czech Republic

E-mail: rahul.sharma@student.upce.cz, {daniel.honc, frantisek.dusek}@upce.cz

KEYWORDS

Mobile robot, dynamic system modelling simulation, trajectory tracking, predictive control.

ABSTRACT

The paper deals with trajectory tracking of the differential drive robot with a mathematical model governing dynamics and kinematics. Motor dynamics and chassis dynamics are considered for deriving a linear state-space dynamic model. Basic nonlinear kinematic equations are linearized into a successively linearized state-space model. The dynamic and kinematic models are augmented to derive a single state-space linear model. The deviation variables are reference variables which are variables of an ideal robot following a reference trajectory which can be pre-calculated. Reference tracking is achieved by model predictive control of supply voltage of both the drive motors by considering constraints on controlled variables and manipulated variables. Simulation results are provided to demonstrate the performance of proposed control strategy in the MATLAB simulation environment.

INTRODUCTION

Trajectory tracking of mobile robots refers to mobile robot tracking in a predefined time-varying reference trajectory, which is one of the fundamental problems in motion control of mobile robots. In the case of differential drive robots, trajectory tracking has been well studied in the past. The most popular way of trajectory tracking is by considering a linearized dynamic error tracking model with feed forward inputs or a successively linearized model.

Model Predictive Control (MPC) is one of the most popular optimization control strategies in the process industries. It is designed to handle complex, constrained, multivariable control problems. It is an online optimization tool, which will generate optimal control actions required at every time instance minimizing an objective function based on predictions (Camacho and Alba 2004). With the increase in computational power, the MPC is not only limited to slow dynamics processes, where dynamical optimization is easily possible, but also there are new applications for faster systems. For

example, MPC control techniques for trajectory tracking of mobile robots as can be seen in (Gu, D. and Hu 2006), (Kuhne et al. 2004) and (Lages et al. 2006). A review of motion control of Wheeled Mobile Robots (WMRs) using MPC can be found in (Kanjawanishkul 2012). A mobile robot trajectory tracking problem with linear and nonlinear state-space MPC is presented in (Kuhne et al. 2005). An experimental overview of WMR is published in (De Luca et al 2001). Dynamic behavior of a differentially steered robot model, where the reference point can be chosen independently and gives us more general formulation, is published in (Dušek et al. 2011). In our previous work (Sharma et al. 2015), we proposed predictive control of the mobile robot, where the linear and angular velocities are optimally controlled by voltages to the drive motors with constraints on controlled variables, manipulated variable and states (current and wheel speed of the motors).

The most common way (for e.g. Maurovic et al. 2011) of trajectory tracking of mobile robots is by controlling the linear and angular velocities by some advanced controllers and then control the mobile robot's wheel speeds by low level controllers like a PID controller. In this paper, we firstly modelled dynamics of the differential drive robot considering motor dynamics and chassis dynamics. The nonlinear kinematics equations are linearized into a linear time-varying error based model by successive linearization, where state variables are deviations from reference variables. Reference variables are variables of the ideal robot which follows a time-varying reference trajectory. The dynamic and kinematic models are augmented into a discrete time-varying state-space model, whose control inputs are motor control variables and outputs are positions in x and y direction and orientation. Model predictive control is used for trajectory tracking simulation in the MATLAB environment by optimizing a quadratic cost function using quadratic programming.

The main advantage of our approach is that (in contrast to the commonly used WMRs models) we consider dynamics of motors as well, so the controller outputs are motor voltages and the robot can be tracked into a reference trajectory, respecting the physical constraints like currents. Since, in trajectory tracking problem, the future set-points are known, MPC is preferred when

compared with other control methods and also because of the ability to handle soft and hard constraints.

MATHEMATICAL MODELLING

The differential drive mobile robot is assumed to have two wheels connected with DC series motors and firmly supported by a castor wheel (See figure 1 and 2).

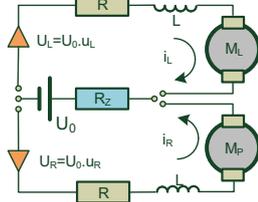


Figure 1: DC Motor Wiring

The mathematical model of the robot, consists of three relatively independent parts. The dynamics of the DC series motor, chassis dynamics (dependency between translational and rotational velocities of the chassis reference point on moments acting to driving wheels), and kinematics (influence of motor speed to translational and rotational velocities).

Dynamics of the mobile robot

The following derivation of the model representing dynamics of the differential drive mobile robot, closely follow the derivations in (Dušek et al. 2011), with some minor notation changes. The Dynamics of the series DC motor can be derived from balancing of voltages (Kirchhoff's law) and balancing of moments. From Kirchhoff's voltage law, we can derive,

$$Ri_L + R_z(i_L + i_R) + L \frac{di_L}{dt} = u_L U_0 - K\omega_L \quad (1)$$

$$Ri_R + R_z(i_L + i_R) + L \frac{di_R}{dt} = u_R U_0 - K\omega_R \quad (2)$$

where, K is the back EMF constant, ω_R and ω_L are the right and left motor speeds. u_R and u_L are the control voltages of the right and left motors respectively. All the other parameters are shown in figure 1.

By considering the balance of moments we can derive,

$$J \frac{d\omega_L}{dt} + k_r \omega_L + M_L = K i_L \quad (3)$$

$$J \frac{d\omega_R}{dt} + k_r \omega_R + M_R = K i_R \quad (4)$$

where J is the moment of inertia of the robot, k_r is the coefficient of rotational resistance. M_L and M_R are the load moments on left and right wheels respectively.

Chassis dynamics is defined with a vector of linear velocity v_B acting on a chassis reference point and with rotation of this vector of angular velocity ω_B (constant for all chassis points). The chassis reference point B is the

point of the intersection of the axis joining the wheels and centre of gravity normal projection – see figure 2. Point T is the general centre of gravity – usually it is placed at the centre of the axis joining the wheels.

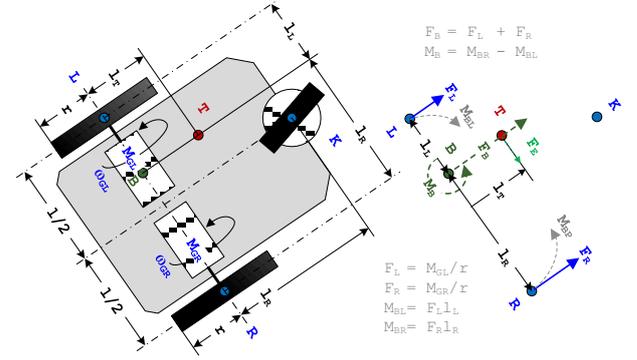


Figure 2: Chassis Scheme and Forces

The chassis dynamics can be expressed by balance of the forces and balance of the moment. Equation (5) is the result of applying balance of forces and Equation (6) from balance of moments.

$$\frac{p_G}{r} M_L + \frac{p_G}{r} M_R - k_v v_B - m \frac{dv_B}{dt} = 0$$

$$M_L + M_R - r_G k_v v_B - r_G m \frac{dv_B}{dt} = 0 \quad (5)$$

$$-l_L \frac{p_G}{r} M_L + l_R \frac{p_G}{r} M_R - k_\omega \omega_B - (J_T + m l_T^2) \frac{d\omega_B}{dt} = 0$$

$$-l_L M_L + l_R M_R - k_\omega \omega_B - r_G J_B \frac{d\omega_B}{dt} = 0 \quad (6)$$

where, p_G is the gear box transmission ratio, k_ω is the resistance coefficient against rotational motion. The rest of the parameters are shown in figure 2. The parameters r_G and J_B are described as,

$$r_G = \frac{r}{p_G} \quad ; \quad J_B = J_T + m l_T^2$$

From the theorems of similar triangles, depicted in figure 3, we can recalculate the peripheral velocities of the wheels v_L , v_R to the linear velocity v_B and angular velocity ω_B at point B as,

$$v_B = \frac{r_G}{l_L + l_R} (l_R \omega_L + l_L \omega_R) \quad (7)$$

$$\omega_B = \frac{r_G}{l_L + l_R} (-\omega_L + \omega_R) \quad (8)$$

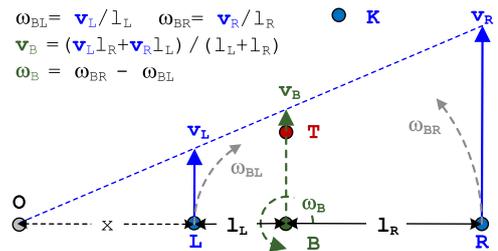


Figure 3: Linear and Angular Velocity Recalculation

These six differential Equations (1)-(6), and two algebraic Equations (7)-(8) containing eight state variables represent a mathematical description of the dynamic behaviour of ideal differentially steered mobile robots with losses linearly dependent on the revolutions or speed. Control signals, u_L and u_R , that control the supply voltages of the motors are input variables.

Calculation of steady-state values for constant engine power voltages are given below. A calculation of steady-state is useful both for the checking of derived equations and for the experimental determination of the values of the unknown parameters. Steady-state in matrix representation is,

$$\begin{bmatrix} R+R_z & R_z & K & 0 & 0 & 0 & 0 & 0 \\ R_z & R+R_z & 0 & K & 0 & 0 & 0 & 0 \\ K & 0 & -k_r & 0 & -1 & 0 & 0 & 0 \\ 0 & K & 0 & -k_r & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -r_G k_v & 0 \\ 0 & 0 & 0 & 0 & -l_L & l_R & 0 & -k_\omega \\ 0 & 0 & l_R & l_L & 0 & 0 & -\frac{l_R+l_L}{r_G} & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & -\frac{l_R+l_L}{r_G} \end{bmatrix} \begin{bmatrix} i_L \\ i_R \\ \omega_L \\ \omega_R \\ M_L \\ M_R \\ v_B \\ \omega_B \end{bmatrix} = \begin{bmatrix} U_L \\ U_R \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (9)$$

The Equation (1-8) can be reduced to a state-space model with four states by introducing the following parameters,

$$\begin{aligned} a_L &= k_r + \frac{k_v l_R r_G^2}{l_L + l_R} & a_R &= k_r + \frac{k_v l_L r_G^2}{l_L + l_R} \\ b_L &= J + \frac{m l_R r_G^2}{l_L + l_R} & b_R &= J + \frac{m l_L r_G^2}{l_L + l_R} \\ c_L &= k_r l_L + \frac{k_\omega r_G^2}{l_L + l_R} & c_R &= k_r l_R + \frac{k_\omega r_G^2}{l_L + l_R} \\ d_L &= J l_L + \frac{J_B r_G^2}{l_L + l_R} & d_R &= J l_R + \frac{J_B r_G^2}{l_L + l_R} \end{aligned}$$

$$\frac{d\mathbf{x}_D}{dt} = \mathbf{A}_D \mathbf{x}_D + \mathbf{B}_D \mathbf{u} \quad (10)$$

$$\mathbf{y}_D = \mathbf{C}_D \mathbf{x}_D$$

where,

$$\mathbf{x}_D = \begin{bmatrix} i_L \\ i_R \\ \omega_L \\ \omega_R \end{bmatrix}; \quad \mathbf{u} = \begin{bmatrix} u_L \\ u_R \end{bmatrix}; \quad \mathbf{y}_D = \begin{bmatrix} v_B \\ \omega_B \end{bmatrix}$$

with matrices \mathbf{A}_D , \mathbf{B}_D and \mathbf{C}_D as,

$$\mathbf{A}_D = \begin{bmatrix} -\frac{R+R_z}{L} & -\frac{R_z}{L} & -\frac{K}{L} & 0 \\ -\frac{R_z}{L} & -\frac{R+R_z}{L} & 0 & -\frac{K}{L} \\ \frac{K(d_R+b_R l_L)}{b_L d_R + b_R d_L} & \frac{K(d_R-b_R l_R)}{b_L d_R + b_R d_L} & -\frac{d_R a_L + b_R c_L}{b_L d_R + b_R d_L} & -\frac{d_R a_R - b_R c_R}{b_L d_R + b_R d_L} \\ \frac{K(d_L-b_L l_L)}{b_L d_R + b_R d_L} & \frac{K(d_L+b_L l_R)}{b_L d_R + b_R d_L} & -\frac{d_L a_L - b_L c_L}{b_L d_R + b_R d_L} & -\frac{d_L a_R + b_L c_R}{b_L d_R + b_R d_L} \end{bmatrix}$$

$$\mathbf{B}_D = \begin{bmatrix} \frac{U_0}{L} & 0 \\ 0 & \frac{U_0}{L} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}; \quad \mathbf{C}_D = \begin{bmatrix} 0 & 0 & \frac{l_R r_G}{l_L + l_R} & \frac{l_L r_G}{l_L + l_R} \\ 0 & 0 & -\frac{r_G}{l_L + l_R} & \frac{r_G}{l_L + l_R} \end{bmatrix}$$

Kinematics of the mobile robot

The following derivations closely follow (Kuhne et al. 2004), despite some notation changes which have been used. Let the global coordinates of the robot be (x_B, y_B) , the orientation of the robot be α , and v_B, ω_B are the linear and angular velocities. The kinematic equations of the differential drive mobile robot is given by (Campion et al. 1996),

$$\begin{aligned} \frac{dx_B}{dt} &= v_B \cos \alpha \\ \frac{dy_B}{dt} &= v_B \sin \alpha \\ \frac{d\alpha}{dt} &= \omega_B \end{aligned} \quad (11)$$

This can be represented as a simple model,

$$\dot{\mathbf{x}}_B = f(\mathbf{x}_B, \mathbf{u}_B) \quad (12)$$

where state variables $\mathbf{x}_B = [x_B \ y_B \ \alpha]^T$ and control inputs $\mathbf{u}_B = [v_B \ \omega_B]^T$.

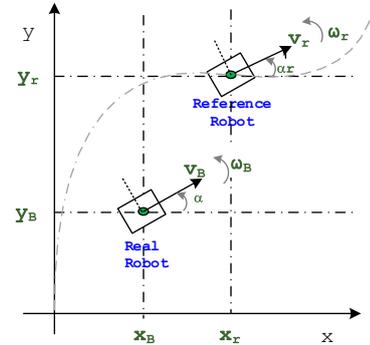


Figure 4: Coordinate System of Real Robot and Reference Robot

A linear model can be derived from the non-linear model, Equations (11), from an error model with respect to the reference robot (see figure 4). A reference robot can be considered as a robot with reference (desired) parameters of the robot to follow a trajectory which can be represented as,

$$\dot{\mathbf{x}}_r = f(\mathbf{x}_r, \mathbf{u}_r) \quad (13)$$

The reference parameters are $[x_r \ y_r \ \alpha_r \ v_r \ \omega_r]$. The linear velocity, orientation angle and angular velocity of the reference robot can be derived from Equation (11) as,

$$v_r(t) = \sqrt{\dot{x}_r(t)^2 + \dot{y}_r(t)^2} \quad (14)$$

$$\alpha_r(t) = \arctan 2(\dot{y}_r(t), \dot{x}_r(t)) \quad (15)$$

$$\omega_r(t) = \dot{\alpha}_r(t) = \frac{\dot{x}_r(t)\ddot{y}_r(t) - \dot{y}_r(t)\ddot{x}_r(t)}{\sqrt{\dot{x}_r(t)^2 + \dot{y}_r(t)^2}} \quad (16)$$

Applying the Taylor series approximation to Equation (12), around the reference points $(\mathbf{x}_r, \mathbf{u}_r)$, we can derive,

$$\begin{aligned} \dot{\mathbf{x}} = & f(\mathbf{x}_r, \mathbf{u}_r) + \left. \frac{\partial f(\mathbf{x}, \mathbf{u})}{\partial \mathbf{x}} \right|_{\substack{\mathbf{x}=\mathbf{x}_r \\ \mathbf{u}=\mathbf{u}_r}} (\mathbf{x}_B - \mathbf{x}_r) + \\ & + \left. \frac{\partial f(\mathbf{x}, \mathbf{u})}{\partial \mathbf{u}} \right|_{\substack{\mathbf{x}=\mathbf{x}_r \\ \mathbf{u}=\mathbf{u}_r}} (\mathbf{u}_B - \mathbf{u}_r) \\ \dot{\mathbf{x}} = & f(\mathbf{x}_r, \mathbf{u}_r) + f_{\mathbf{x}_r}(\mathbf{x}_B - \mathbf{x}_r) + f_{\mathbf{u}_r}(\mathbf{u}_B - \mathbf{u}_r) \end{aligned} \quad (17)$$

Subtracting Equation (17) from Equation (13) gives,

$$\dot{\bar{\mathbf{x}}} = f_{\mathbf{x}_r} \bar{\mathbf{x}} + f_{\mathbf{u}_r} \bar{\mathbf{u}} \quad (18)$$

$\bar{\mathbf{x}}$ is the error vector of state variables and $\bar{\mathbf{u}}$ is the error vector of control variables with respect to the reference robot. The approximation of $\dot{\bar{\mathbf{x}}}$ in Equation (18), by the forward differences gives the following discrete-time linear time-variant (LTV) state-space model:

$$\begin{aligned} \bar{\mathbf{x}}_{\mathbf{K}}(k+1) = & \bar{\mathbf{A}}_{\mathbf{K}}(k)\bar{\mathbf{x}}_{\mathbf{K}}(k) + \bar{\mathbf{B}}_{\mathbf{K}}(k)\bar{\mathbf{u}}_{\mathbf{K}}(k) \\ \bar{\mathbf{y}}_{\mathbf{K}}(k) = & \bar{\mathbf{C}}_{\mathbf{K}}\bar{\mathbf{x}}_{\mathbf{K}}(k) \end{aligned} \quad (19)$$

$$\begin{aligned} \bar{\mathbf{A}}_{\mathbf{K}}(k) = & \begin{bmatrix} 1 & 0 & -v_r(k)\sin\alpha_r(k)T \\ 0 & 1 & v_r(k)\cos\alpha_r(k)T \\ 0 & 0 & 1 \end{bmatrix} \\ \bar{\mathbf{B}}_{\mathbf{K}}(k) = & \begin{bmatrix} \cos\alpha_r(k)T & 0 \\ \sin\alpha_r(k)T & 0 \\ 0 & T \end{bmatrix}; \bar{\mathbf{C}}_{\mathbf{K}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \bar{\mathbf{x}}_{\mathbf{K}} = & \begin{bmatrix} x_B(k) - x_r(k) \\ y_B(k) - y_r(k) \\ \alpha(k) - \alpha_r(k) \end{bmatrix}; \bar{\mathbf{u}}_{\mathbf{K}} = \begin{bmatrix} v_B(k) - v_r(k) \\ \omega_B(k) - \omega_r(k) \end{bmatrix} \end{aligned}$$

where T is the sampling period and $\bar{\mathbf{x}}_{\mathbf{K}}$ is deviation state vector which represents the error with respect to the reference robot, and $\bar{\mathbf{u}}_{\mathbf{K}}$ is associated with the control input. The reference values, v_r, α_r, ω_r are the reference linear velocity, orientation angle and angular velocity respectively which can be calculated from Equations (14-16).

Combined model – LTV

The kinematic model is linearized into a discrete error model. The dynamic model also has to be converted to a discrete error based model for augmenting with a kinematic model. Since the dynamic model is linear time

invariant, the error model will be the same as that of Equation (10) but has to be discretized. Let the following be the discrete time state-space dynamic model.

$$\begin{aligned} \bar{\mathbf{x}}_{\mathbf{D}}(k+1) = & \bar{\mathbf{A}}_{\mathbf{D}}(k)\bar{\mathbf{x}}_{\mathbf{D}}(k) + \bar{\mathbf{B}}_{\mathbf{D}}(k)\bar{\mathbf{u}}_{\mathbf{D}}(k) \\ \bar{\mathbf{y}}_{\mathbf{D}}(k) = & \bar{\mathbf{C}}_{\mathbf{D}}(k)\bar{\mathbf{x}}_{\mathbf{D}}(k) \end{aligned}$$

The matrices $\bar{\mathbf{A}}_{\mathbf{D}}, \bar{\mathbf{B}}_{\mathbf{D}}$ and $\bar{\mathbf{C}}_{\mathbf{D}}$ are discretized matrices of the dynamic model (Equation (10)). The state variables and control inputs are deviation variables from the reference points, $\mathbf{x}_{\mathbf{D}_r}$ and $\mathbf{u}_{\mathbf{D}_r}$, as,

$$\bar{\mathbf{x}}_{\mathbf{D}}(k) = \begin{bmatrix} i_L(k) - i_{L_r}(k) \\ i_R(k) - i_{R_r}(k) \\ \omega_L(k) - \omega_{L_r}(k) \\ \omega_R(k) - \omega_{R_r}(k) \end{bmatrix}; \bar{\mathbf{u}}_{\mathbf{D}}(k) = \begin{bmatrix} u_L(k) - u_{L_r}(k) \\ u_R(k) - u_{R_r}(k) \end{bmatrix}$$

This dynamics model and linearized kinematic time-variant model, Equation (19), can be augmented into a single state-space time-variant model with 9 states (currents, wheel speeds, linear and angular velocities and coordinates), two control variables (motor voltage control input) and three outputs (position in x and y direction and orientation measured from x direction).

$$\begin{aligned} \bar{\mathbf{x}}(k+1) = & \bar{\mathbf{A}}(k)\bar{\mathbf{x}}(k+1) + \bar{\mathbf{B}}(k)\bar{\mathbf{u}}(k) \\ \bar{\mathbf{y}}(k) = & \bar{\mathbf{C}}(k)\bar{\mathbf{x}}(k) \end{aligned} \quad (20)$$

where,

$$\begin{aligned} \bar{\mathbf{A}}(k) = & \begin{bmatrix} \bar{\mathbf{A}}_{\mathbf{D}(4 \times 4)} & \mathbf{0}_{(4 \times 5)} \\ \bar{\mathbf{C}}_{\mathbf{D}(2 \times 4)} & \mathbf{0}_{(2 \times 5)} \\ \mathbf{0}_{(3 \times 4)} & \bar{\mathbf{B}}_{\mathbf{K}(3 \times 2)} & \bar{\mathbf{A}}_{\mathbf{K}(3 \times 3)} \end{bmatrix}; \bar{\mathbf{B}}(k) = \begin{bmatrix} \bar{\mathbf{B}}_{\mathbf{D}(4 \times 2)} \\ \mathbf{0}_{(5 \times 2)} \end{bmatrix} \\ \bar{\mathbf{C}}(k) = & \begin{bmatrix} \mathbf{0}_{(3 \times 6)} & \bar{\mathbf{C}}_{\mathbf{K}(3 \times 3)} \end{bmatrix}; \bar{\mathbf{x}} = [\bar{\mathbf{x}}_{\mathbf{D}} \ \bar{\mathbf{u}}_{\mathbf{K}} \ \bar{\mathbf{x}}_{\mathbf{K}}]^T; \bar{\mathbf{u}} = \bar{\mathbf{u}}_{\mathbf{D}} \end{aligned}$$

MODEL PREDICTIVE CONTROL

At each sampling time, the model predictive controller generates an optimal control sequence by optimizing a quadratic cost function. The first control action of this sequence is applied to the system. The optimization problem is solved again at the next sampling time using the updated process measurements and a shifted horizon. The cost function formulation depends on the control requirements. The most common cost function is in the form of,

$$\begin{aligned} J = & \sum_{i=1}^{n_y} \sum_{j=1}^{N_2} r_j [\hat{y}_i(k+j) - w_i(k+j)]^2 + \\ & + \sum_{i=1}^{n_u} \sum_{j=1}^{N_3} q_j [\Delta u_i(k+j-1)]^2 \end{aligned} \quad (21)$$

Where $\hat{y}_i(k+j)$ is an optimum j -step ahead prediction of the system i -th output, N_2 is the control error horizon,

N_3 is the control horizon and $w_i(k+j)$ is the future set-point or reference for the i -th controlled variable. The parameters, r_i and q_i are the weighting coefficient for control errors and control increments respectively. $\Delta u_i(k+j-1)$ is the control increment of the i -th input. n_u and n_y are the number of inputs and number of outputs (manipulated and controlled variables).

The cost function consists of two parts, mainly costs due to control error during the control error horizon N_2 and costs to penalize the control signal increments during the control horizon N_3 . For simplicity in the following text we consider, $N_2=N_3=N$.

A general discrete-time state-space model is given as,

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) \end{aligned} \quad (22)$$

An incremental state-space model can also be used, if the model input is the control increment $\Delta\mathbf{u}(k)$ instead of $\mathbf{u}(k)$. $\Delta\mathbf{u}(k)=\mathbf{u}(k)-\mathbf{u}(k-1)$

$$\begin{aligned} \begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{u}(k) \end{bmatrix}_{\mathbf{x}_p(k+1)} &= \underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{u}(k-1) \end{bmatrix}_{\mathbf{x}_p(k)} + \underbrace{\begin{bmatrix} \mathbf{B} \\ \mathbf{I} \end{bmatrix}}_{\mathbf{P}} \Delta\mathbf{u}(k) \\ \mathbf{y}(k) &= \underbrace{\begin{bmatrix} \mathbf{C} & \mathbf{0} \end{bmatrix}}_{\mathbf{O}} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{u}(k-1) \end{bmatrix}_{\mathbf{x}_p(k)} \end{aligned} \quad (23)$$

The predicted output representation of state-space model, in matrix form, is

$$\begin{aligned} \underbrace{\begin{bmatrix} \hat{\mathbf{y}}(k+1) \\ \hat{\mathbf{y}}(k+2) \\ \vdots \\ \hat{\mathbf{y}}(k+N) \end{bmatrix}}_{\mathbf{Y}} &= \underbrace{\begin{bmatrix} \mathbf{OP} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{OMP} & \mathbf{OP} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{OM}^{N-1}\mathbf{P} & \mathbf{OM}^{N-2}\mathbf{P} & \dots & \mathbf{OP} \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} \Delta\mathbf{u}(k) \\ \Delta\mathbf{u}(k+2) \\ \vdots \\ \Delta\mathbf{u}(k+N-1) \end{bmatrix}}_{\mathbf{U}} + \underbrace{\begin{bmatrix} \mathbf{OM} \\ \mathbf{OM}^2 \\ \vdots \\ \mathbf{OM}^N \end{bmatrix}}_{\mathbf{f}} \mathbf{x}_p \end{aligned}$$

Which can be represented as sum of forced and free responses,

$$\mathbf{Y} = \underbrace{\mathbf{GU}}_{\text{forced response}} + \underbrace{\mathbf{f}}_{\text{free response}} \quad (24)$$

Cost function

The cost function in Equation (21) can be represented in matrix format as,

$$\mathbf{J} = (\mathbf{Y} - \mathbf{W})^T \mathbf{R}(\mathbf{Y} - \mathbf{W}) + \mathbf{U}^T \mathbf{Q}\mathbf{U} \quad (25)$$

where, \mathbf{R} and \mathbf{Q} are diagonal matrices with diagonal elements r_i and q_i respectively and \mathbf{W} is a column vector of N future set points.

Constraints

In a long range predictive control, the controller has to anticipate constraint violation and correct control actions in an appropriate way. The input constraints are,

$$\begin{aligned} \mathbf{u}_{min} \leq \mathbf{u}(i) \leq \mathbf{u}_{max}, \quad i \in \{k, k+N-1\} \\ \mathbf{y}_{min} \leq \mathbf{y}(i) \leq \mathbf{y}_{max}, \quad i \in \{k+1, k+N\} \end{aligned} \quad (26)$$

The implementation of MPC with constraints involves the minimization of a quadratic cost function

$$J = \mathbf{U}^T \underbrace{(\mathbf{G}^T \mathbf{R} \mathbf{G} + \mathbf{Q})}_{\mathbf{H}} \mathbf{U} + 2 \underbrace{(\mathbf{f} - \mathbf{W})^T \mathbf{R} \mathbf{G}}_{\mathbf{g}'} \mathbf{U} + \underbrace{(\mathbf{f} - \mathbf{W})^T \mathbf{R} (\mathbf{f} - \mathbf{W})}_{\mathbf{k}}$$

Subject to the linear inequalities $\mathbf{A}\mathbf{U} \leq \mathbf{b}$, which is a Quadratic Programming (QP) problem. The QP problem can be solved e.g. using the function *quadprog* in MATLAB (Honc and Dušek 2013).

Predictive control of mobile robot

The augmented model is an error based model whose state variables are deviations from reference variables. The reference variables can be seen as an ideal robot following a time-varying reference trajectory. These reference velocities v_r , ω_r and orientation angle α_r can be calculated from Equations (14) to (16) from the reference inputs (positional coordinates of the robot - x_r , y_r). The other reference variables $\mathbf{x}_{D,r}$ and $\mathbf{u}_{D,r}$ can be pre-calculated from the model, Equation (10), by closed loop control with set-points (as previously calculated) v_r , ω_r and with an initial condition calculated from steady state Equation (9). The trajectory tracking of the mobile robot is achieved by model predictive control with the linear time-variant model, Equation (20), with a cost function as in Equation (25) considering the constraints, Equation (26). At every time instance, the MPC algorithm will calculate the optimal control inputs (motor voltage control inputs - u_L and u_R). The overall control scheme is shown in figure 5.

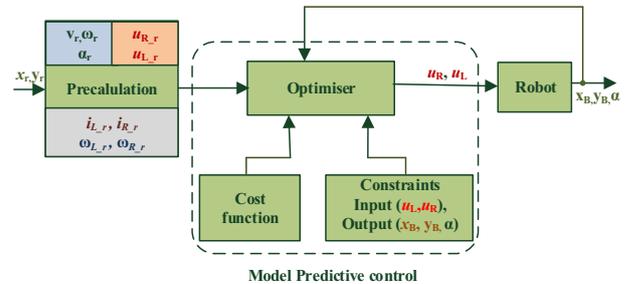


Figure 5: Overall Control Scheme

SIMULATION RESULTS

Chassis parameters and DC motor parameters were chosen as in (Dušek et al. 2011). These values are chosen so that they roughly correspond to the real physical values of the mobile robot. The reference trajectory chosen was an S-shaped trajectory as follows,

$$\begin{aligned} x_r &= \sin(2t) \\ y_r &= 5 \sin\left(t + \frac{\pi}{2}\right) \end{aligned}$$

The mobile robot MPC was simulated in the MATLAB simulation environment with a sample time of 0.1 s and prediction horizon, $N=5$. The initial position of the robot was chosen to be the same as the reference trajectory points. The weighing matrices are chosen as,

$$\begin{aligned} \mathbf{Q} &= \text{diag}(1, 10, 10) \\ \mathbf{R} &= \text{diag}(10, 10) \end{aligned}$$

In figure 6, the simulated trajectory is compared with the desired (reference trajectory). The control inputs and reference (calculated by Equation 14 and 16) and simulated linear and angular velocities are shown in figure 7. Figure 8 depicts the wheel speeds and currents. Figure 9 shows the reference orientation (calculated by Equation (15)) and simulated orientation.

Constraints were applied to controlled variables (control voltages to right and left wheel). The constraints of control voltage of the motors were set to $[0, 1]$ since the source voltage is 10 V and no backward motions of motor was assumed. The trajectory was chosen in such a way that we can see the response of the robot when a sudden change of position and orientation to the robot.

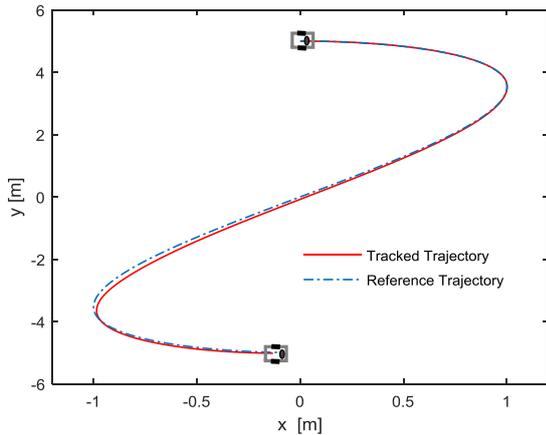


Figure 6: Trajectory Tracking

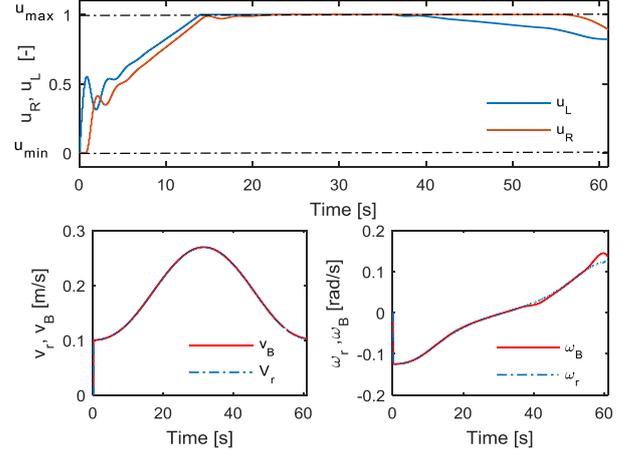


Figure 7: Control Inputs, Linear and Angular Velocity

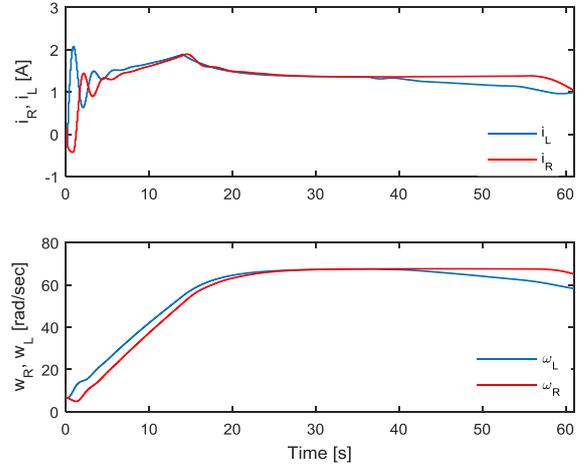


Figure 8: Currents and Wheel Speeds

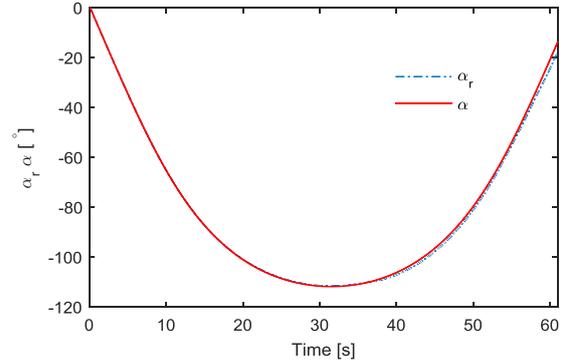


Figure 9: Reference and Simulated Orientation

Since the main objective of the paper was to model and simulate the response, efforts were not made in the control quality (e.g. constraints on state variables, tuning of weighing matrices, steady state error etc.). Control quality can be significantly improved by proper tuning of weighing matrices and/or by choosing an optimal horizon and/or by including a state observer etc.

CONCLUSION

In this paper, a linear time-variant model is derived by considering both the kinematics and dynamics of the mobile robot, which will allow trajectory tracking of mobile robot by controlling the control voltage to the motors. Constraints were considered only for the control variable.

As a future research direction, we are looking to incorporate other issues into our MPC formulation, such as including constraints on wheel speeds and currents, decreasing the computation time etc. Moreover, we expect to finish this controller implementation in a real robot and to conduct real experiments with the mobile robot in various environments. Path planning, obstacle avoidance etc., are other elements we wish to consider.

This research was supported by project SGS_2016_021, Mobile Robot Motion Control with Model Predictive Controller at FEL, University of Pardubice. This support is very gratefully acknowledged.

REFERENCES

- Camacho, E.F. and Alba, C.B., 2013. *Model predictive control*. Springer Science & Business Media.
- Gu, D. and Hu, H., 2006. Receding horizon tracking control of wheeled mobile robots. *Control Systems Technology, IEEE Transactions on*, 14(4), pp.743-749.
- Kuhne, F., Lages, W.F. and da Silva Jr, J.M.G., 2004. Model predictive control of a mobile robot using linearization. In *Proceedings of mechatronics and robotics* (pp. 525-530).
- Lages, W.F. and Alves, J.A.V., 2006, September. Real-time control of a mobile robot using linearized model predictive control. In *Proc. of 4th IFAC Symposium on Mechatronic Systems* (pp. 968-973).
- Kanjanawanishkul, K., 2012. Motion control of a wheeled mobile robot using model predictive control: A survey. *KKU Research Journal*, 17(5), pp.811-837.
- Künhe, F., Gomes, J. and Fetter, W., 2005, September. Mobile robot trajectory tracking using model predictive control. In *II IEEE latin-american robotics symposium*.
- De Luca, A., Oriolo, G. and Vendittelli, M., 2001. Control of wheeled mobile robots: An experimental overview. In *Ramsete* (pp. 181-226). Springer Berlin Heidelberg.
- Dušek, F., Honc, D. and Rozsival, P., 2011. Mathematical model of differentially steered mobile robot. In *18th International Conference on Process Control, Tatranská Lomnica, Slovakia*.
- Honc, D. and Dusek, F., 2013. State-Space Constrained Model Predictive Control. In *ECMS* (pp. 441-445).
- Sharma, K.R., Honc, D. and Dušek, F., 2015. Model Predictive Control of Trajectory Tracking of Differentially Steered Mobile Robot. In *Intelligent Data Analysis and Applications* (pp. 85-95). Springer International Publishing.
- Campion, G., Bastin, G. e D'Andréa-Novel B., 1996. Structural properties and classification of kinematic and dynamical models of wheeled mobile robots, *IEEE Transactions on Robotics and Automation* 12 (1): pp.47-62
- Maurovic, I., Baotic, M. and Petrovic, I., 2011, July. Explicit model predictive control for trajectory tracking with mobile robots. In *Advanced Intelligent Mechatronics (AIM), 2011 IEEE/ASME International Conference on* (pp. 712-717). IEEE.

AUTHOR BIOGRAPHIES



RAHUL SHARMA K., was born in Kochi, India and went to the Amrita University, where he studied electrical engineering and obtained his M.Tech degree in 2013. He is now doing his Ph.D. studies at the Department of process control, Faculty of Electrical and Informatics, University of Pardubice, Czech Republic.

e-mail: rahul.sharma@student.upce.cz



DANIEL HONC was born in Pardubice, Czech Republic and studied at the University of Pardubice in the field of Process Control and obtained his Ph.D. degree in 2002. He is the head of the Department of Process Control at the Faculty of Electrical Engineering and Informatics.

e-mail: daniel.honc@upce.cz



FRANTIŠEK DUŠEK was born in Dačice, Czech Republic and studied at the Pardubice Faculty of Chemical Technology in the field of Automation and obtained his MSc. degree in 1980. He worked for the pulp and paper research institute IRAPA. Now he is the vice-dean of the Faculty of Electrical Engineering and Informatics. In 2001 he became an Associate Professor.

e-mail: frantisek.dusek@upce.cz

PREDICTIVE AND FEEDBACK LINEARIZING CONTROL OF *CHLAMYDOMONAS REINHARDTII* PHOTOAUTOTROPHIC GROWTH PROCESS

Florin Stîngă and Emil Petre
Department of Automation and Electronics
Faculty of Automation, Computers and Electronics
University of Craiova
Bd. Decebal, no. 107, Craiova, Romania
E-mail: [florin, epetre]@automation.ucv.ro

INTRODUCTION

It is well known that water is one of the essential elements of life being an important resource both for industrial applications and domestic usage. Lately, many environmental laws and directives have been enforced in order to decrease the industrial and urban pollution. This situation has led to an increase in the use of wastewater biological treatment processes using anaerobic digestion. This fermentation bioprocess is very important since it produces valuable energy (methane) besides removing the organic pollution from the liquid influent. It is useful for concentrated wastes such as agricultural and food industry wastewater (F. Angulo et al. 2007). Nevertheless, its main drawback is the production of carbon dioxide (CO₂) and its easy destabilization, giving rise to the disappearance of the methanogenic bacteria (G. Bastin and D. Dochain 1990; O. Bernard 2004). Therefore in the last decade the researchers have been looking for solutions to improve the efficiency in the pollution reduction and for CO₂ mitigation (O. Bernard 2011; G.A. Ifrim 2012; G.A. Ifrim et al. 2013, S. Tebbani et al. 2014). A recently used solution consist in the growth of some microalgae populations that by using light as source of energy are able to assimilate inorganic forms of carbon (CO₂, HCO₃⁻) and to convert them into requisite organic substances for cellular functions, generating at the same time oxygen O₂ (G.A. Ifrim 2012; G.A. Ifrim et al. 2013, S. Tebbani et al. 2014; S. Tebbani et al. 2013; S. Tebbani et al. 2015).

The control of such processes remains a key issue for the improvement of stability and process efficiency. A difficulty for the design of high-performance control techniques of such living processes lies in the fact that, in many cases, the models contain kinetic parameters and/or yield coefficients that are highly uncertain and time varying (G. Bastin and D. Dochain 1990; O. Bernard 2004; D. Dochain and P. Vanrolleghem 2001; D.J. Batstone et. Al 2002; F. Mairet et al. 2011; F. Mairet et al. 2011a). Another problem in the control of these processes is finding adequate sensors used for measuring all the important state variables (G. Bastin and D. Dochain 1990; O. Bernard 2011). The problem becomes of great importance in complex systems like wastewater treatment plants, where critical instability of the process must be avoided, making the monitoring of the system

variables an important issue (F. Angulo et al. 2007; G.A. Ifrim et al. 2013; D. Dochain 2008; E. Petre et al. 2013). To surmount these difficulties, numerous control strategies were developed such as linearizing feedback (F. Angulo et al. 2007; G.A. Ifrim et al. 2013; I. Neria-González et al. 2009), adaptive and robust-adaptive approach (G. Bastin and D. Dochain), predictive an optimal control (F. Logist 2011; S. Tebbani et al. 2014), sliding mode (D. Selişteanu et al. 2007), and so on. This paper presents the design of predictive and feedback linearizing control methods for a complex photoautotrophic growth process that is carried out in a torus photobioreactor numerical simulation are performed in order to validate the proposed control algorithms.

MATHEMATICAL MODEL

The considered bioprocess is a photoautotrophic growth of the green alga *C. reinhardtii* in a photobioreactor under limiting conditions. Microalgae are able to absorb CO₂ as major substrate and to generate O₂ as residue from the water oxidation reaction induced by the light as source of energy (G.A. Ifrim et al. 2013; S. Tebbani et al. 2013). A dynamical model employed for describing all phenomena that is carried out in the photobioreactor, assuming well mixed conditions, was developed in (G.A. Ifrim 2012). In this paper a simplified model, described by the following differential equations, is used:

$$\begin{aligned} \dot{X}(t) &= \langle \tau_X \rangle - D(t)X(t) \\ \dot{C}_{TIC}(t) &= -\langle \tau_{TIC} \rangle + N_{CO_2}(t) + D(t) \begin{pmatrix} C_{TIC,i}^- \\ -C_{TIC}(t) \end{pmatrix} \\ \dot{C}_{O_2}(t) &= \langle \tau_{O_2} \rangle + N_{O_2}(t) - D(t)C_{O_2}(t) \\ \dot{y}_{out}^{CO_2}(t) &= \frac{RT_a}{PV_g} \left(-y_{out}^{CO_2}(t)G_{out}(t) - V_l N_{CO_2}(t) + G_{in}^{CO_2}(t) \right) \\ \dot{y}_{out}^{O_2}(t) &= \frac{RT_a}{PV_g} \left(-y_{out}^{O_2}(t)G_{out}(t) - V_l N_{O_2}(t) + G_{in}^{O_2}(t) \right) \end{aligned} \quad (1)$$

where X is the biomass concentration, C_{TIC} is total inorganic carbon concentration, and C_{O_2} is the dissolved oxygen concentration, $y_{out}^{CO_2}$ and $y_{out}^{O_2}$ are the molar fractions of CO_2 and O_2 in the outlet gas, D is the dilution rate, $C_{TIC,i}$ is the concentrations of TIC in the feed, R is the universal gas constant, V_g and V_l are the gas and liquid volume of the photobioreactor, P is the total pressure in the gas phase, T_a is the temperature, $G_{in}^{CO_2}$ and $G_{in}^{O_2}$ are the CO_2 and O_2 feeding flowrates, respectively, and G_{out} are the total output flowrate. The global volumetric growth rate τ_x is expressed using a Monod model of the light flux into the culture medium as:

$$\langle \tau_x \rangle = \left(\frac{\mu_{\max}}{L} \int_0^L \frac{I(z)}{K_i + I(z)} dz - \mu_s \right) X, \quad (2)$$

where μ_{\max} is the maximum specific growth rate, K_i is the half-saturation constant, L is the photobioreactor total depth and $I(z)$ is the local value of the irradiance, expressed through a radiative model for a specific photobioreactor. In case of torus configuration the expression of $I(z)$ is given by (G.A. Ifrim et al. 2013):

$$I(z) = 2q_0 \frac{(1+\alpha)e^{\delta(L-z)} - (1-\alpha)e^{-\delta(L-z)}}{(1+\alpha)^2 e^{\delta L} - (1-\alpha)^2 e^{-\delta L}}, \quad (3)$$

where q_0 represents the hemispherical incident light flux, $\delta = X\sqrt{E_a(E_a + 2bE_s)}$ is the two-flux extinction coefficient, $\alpha = \sqrt{E_a/(E_a + 2bE_s)}$ is the linear scattering modulus, E_a is the mass absorption coefficient, E_s is the mass scattering coefficient and b is the backward scattering fraction. The microalgae growth is synthesized exclusively from inorganic carbon, so the TIC consumption rate is given by:

$$\langle \tau_{TIC} \rangle = \frac{1}{M_x} \langle \tau_x \rangle, \quad (4)$$

where M_x is the C-mole mass of the cells and $\langle \rangle$ denotes spatial averaging. The oxygen production rates is proportional to the growth rate, and can be expressed as:

$$\langle \tau_{O_2} \rangle = \frac{Q_p}{M_x} \langle \tau_x \rangle, \quad (5)$$

where Q_p is the photosynthetic quotient.

The CO_2 and O_2 mass transfer rates to the liquid phase are expressed as follows (S. Tebbani et al. 2013):

$$N_{CO_2} = (K_L a)_{CO_2} \left(\frac{P}{H_{CO_2}} y_{out}^{CO_2} - C_{CO_2} \right), \quad (6)$$

$$N_{O_2} = (K_L a)_{O_2} \left(\frac{P}{H_{O_2}} y_{out}^{O_2} - C_{O_2} \right), \quad (7)$$

where $(K_L a)_{CO_2}$ and $(K_L a)_{O_2}$ are volumetric mass transfer coefficients, H_{CO_2} and H_{O_2} are the Henry constant at 25°C. Also, the CO_2 concentrations is given by (S. Tebbani et al. 2013) :

$$C_{CO_2} = \frac{C_{TIC}}{(1 + K_1 10^{pH} + K_1 K_2 10^{2pH})}. \quad (8)$$

Finally, the output flow rate of gas mixtures is given by:

$$G_{out} = \frac{G_{in}^{N_2}}{1 - y_{out}^{CO_2} - y_{out}^{O_2}}. \quad (9)$$

So, the above model (1) is nonlinear and control-affine model with respect to the state variables (S. Tebbani et al. 2015).

The control objective of this study is to maintain a biomass setpoint concentration in the photobioreactor and to minimize the quantity of CO_2 in the output flux, using as manipulated variables the dilution rate (D) and the feeding flowrate of CO_2 ($G_{in}^{CO_2}$).

In order to obtain a less complex model, the global volumetric growth rate is approximated by the following expression:

$$\langle \tau_x \rangle = (c_1 e^{c_2 X} - \mu_s) X, \quad (10)$$

where c_1 and c_2 are constants, defined in accordance with μ_{\max} , L , α , q_0 , K_I and $\delta_1 = \sqrt{E_a(E_a + 2bE_s)}$.

Also, it is considered that the feeding flowrate of the oxygen is given by:

$$G_{in}^{O_2} = c_3 G_{in}^{CO_2} \quad (11)$$

where c_3 is a subunitary constant.

In conclusion, the model of microalgae growth can be cast in general form:

$$\begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= h(x) \end{aligned} \quad (12)$$

where $u = [u_1 \quad u_2]^T = [D \quad G_{in}^{CO_2}]^T$

$$x = [x_1 \ x_2 \ x_3 \ x_4 \ x_5]^T = [X \ C_{TIC} \ C_{O_2} \ y_{out}^{CO_2} \ y_{out}^{O_2}]^T,$$

$$f(x) = \begin{bmatrix} \langle \tau_x \rangle \\ -\langle \tau_{TIC} \rangle + N_{CO_2} \\ \langle \tau_{O_2} \rangle + N_{O_2} \\ -G_{out} - V_l N_{CO_2} \\ -G_{out} - V_l N_{O_2} \end{bmatrix}, g(x) = \begin{bmatrix} -x_1 & 0 \\ C_{TIC,i} - x_2 & 0 \\ -x_3 & 0 \\ 0 & \frac{RT_a}{PV_g} \\ 0 & \frac{RT_a c_3}{PV_g} \end{bmatrix},$$

$$h(x) = \begin{bmatrix} x_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_4 & 0 \end{bmatrix},$$

Also, the variables $\langle \tau_x \rangle$, $\langle \tau_{TIC} \rangle$, $\langle \tau_{O_2} \rangle$, N_{CO_2} , N_{O_2} and G_{out} are given by relations (10), (4), (5), (6), (7) and (9).

PREDICTIVE CONTROL

The model predictive control uses a discrete model of the system in order to obtain a prediction of its future behaviour, by applying a set of input sequence to model, taking into account constraints on state, output and input variables (M. Morari and J.H. Lee 1999). The linearizing model is defined around certain equilibrium points, being obtained by Jacobian linearization and Taylor series expansion, respectively,

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + B\tilde{u}(t) \\ \tilde{y}(t) &= C\tilde{x}(t) \end{aligned} \quad (13)$$

where:

$$\begin{aligned} \tilde{x}(t) &= [\tilde{x}_1(t) \ \tilde{x}_2(t) \ \tilde{x}_3(t) \ \tilde{x}_4(t) \ \tilde{x}_5(t)]^T, \\ \tilde{u}(t) &= [\tilde{u}_1 \ \tilde{u}_2]^T \\ \tilde{x}_i(t) &= x_i(t) - \bar{x}_i, \quad i = 1 \dots 5 \\ \tilde{u}_j(t) &= u_j(t) - \bar{u}_j, \quad j = 1, 2 \end{aligned}$$

\bar{x}_i and \bar{u}_j are equilibrium points,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix},$$

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \\ b_{51} & b_{52} \end{bmatrix}, C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$a_{11} = c_1 e^{c_2 \bar{x}_1} - \bar{u}_1 - \mu_s + c_1 c_2 \bar{x}_1 e^{c_2 \bar{x}_1}, a_{12} = 0, a_{13} = 0, a_{14} = 0, a_{15} = 0,$$

$$a_{21} = \left(\frac{1}{M_x} \right) (\mu_s - c_1 e^{c_2 \bar{x}_1} - c_1 c_2 \bar{x}_1 e^{c_2 \bar{x}_1}),$$

$$a_{22} = -\bar{u}_1 - \frac{(K_L a)_{CO_2}}{K_1 10^{pH} + K_1 K_2 10^{2pH} + 1}, a_{23} = 0,$$

$$a_{24} = \frac{(K_L a)_{CO_2}}{H_{CO_2}}, a_{25} = 0,$$

$$a_{31} = \left(\frac{Q_p}{M_x} \right) (c_1 c_2 \bar{x}_1 e^{c_2 \bar{x}_1} - \mu_s + c_1 e^{c_2 \bar{x}_1}),$$

$$a_{32} = 0, a_{33} = -(K_L a)_{O_2} - \bar{u}_1, a_{34} = 0, a_{35} = \frac{(K_L a)_{O_2} P}{H_{O_2}},$$

$$a_{41} = 0,$$

$$a_{42} = \frac{(K_L a)_{CO_2} RT_a V_l}{PV_g K_1 10^{pH} + K_1 K_2 10^{2pH} + 1}, a_{43} = 0,$$

$$a_{44} = -\frac{RT_a}{PV_g} \left(\frac{G_{in}^{N_2} \bar{x}_4}{(\bar{x}_4 + \bar{x}_5 - 1)^2} - \frac{G_{in}^{N_2}}{(\bar{x}_4 + \bar{x}_5 - 1)} + \frac{(K_L a)_{CO_2} PV_l}{H_{CO_2}} \right),$$

$$a_{45} = -\frac{G_{in}^{N_2} RT_a \bar{x}_4}{PV_g (\bar{x}_4 + \bar{x}_5 - 1)^2}, a_{51} = 0, a_{52} = 0,$$

$$a_{53} = \frac{(K_L a)_{O_2} RT_a V_l}{PV_g}, a_{54} = -\frac{G_{in}^{N_2} RT_a \bar{x}_5}{PV_g (\bar{x}_4 + \bar{x}_5 - 1)^2},$$

$$a_{55} = -\frac{RT_a}{PV_g} \left(\frac{G_{in}^{N_2} \bar{x}_5}{(\bar{x}_4 + \bar{x}_5 - 1)^2} - \frac{G_{in}^{N_2}}{(\bar{x}_4 + \bar{x}_5 - 1)} + \frac{(K_L a)_{O_2} PV_l}{H_{O_2}} \right),$$

$$b_{11} = -\bar{x}_1, b_{12} = 0, b_{21} = -C_{TIC,i} - \bar{x}_2, b_{22} = 0,$$

$$b_{31} = -\bar{x}_3, b_{32} = 0, b_{41} = 0, b_{42} = \frac{RT_a}{PV_g}, b_{51} = 0,$$

$$b_{52} = \frac{RT_a c_3}{PV_g}.$$

In order to obtain the predictions of the future behaviour of the system, a discrete Euler approximation of the above linear system, was used,

$$\begin{aligned} \tilde{x}(k+1) &= \underbrace{(I + AT_e)}_{A_d} \tilde{x}(k) + \underbrace{(BT_e)}_{B_d} \tilde{u}(k) \\ \tilde{y}(k) &= C_d \tilde{x}(k) \end{aligned} \quad (14)$$

where: $\dot{\tilde{x}}(t)|_{t=kT_e} \approx \frac{\tilde{x}(k+1) - \tilde{x}(k)}{T_e}$, T_e is the sampling time, $\tilde{x}(kT_e) = \tilde{x}(k)$ and $C_d = C$.

Additionally, an integral action has been embedded in the closed loop behavior of the system, such that:

$$\begin{aligned} x_a(k+1) &= A_a x_a(k) + B_a \Delta u_a(k) \\ y_a(k) &= C_a x_a(k) \end{aligned} \quad (15)$$

where:

$$\begin{aligned} x_a(k) &= \begin{bmatrix} \Delta \tilde{x}(k) & \tilde{y}(k) \end{bmatrix}^T, \\ \Delta \tilde{x}(k) &= \tilde{x}(k) - \tilde{x}(k-1), A_a = \begin{bmatrix} A_d & o_{p \times n}^T \\ C_d A_d & I_{p \times p} \end{bmatrix}, \\ B_a &= \begin{bmatrix} B_d \\ C_d B_d \end{bmatrix}, \Delta u_a(k) = \tilde{u}(k) - \tilde{u}(k-1), \\ C_a &= \begin{bmatrix} o_{p \times n} & I_{p \times p} \end{bmatrix}, o_{p \times n} \text{ is a zero matrix, } I_{p \times p} \text{ is the} \\ &\text{identity matrix, } p \text{ is the number of outputs and } n \text{ is the} \\ &\text{number of states.} \end{aligned}$$

According with the considered outputs, the control objective is to maintain the biomass concentration and the output molar fraction of the CO₂ at certain set points. The optimal commands, represented by the dilution rate D and the feeding CO₂ flowrate $G_{in}^{CO_2}$ are computed at every step by solving the quadratic minimization problem and applying the principle of receding horizon strategy:

$$\begin{aligned} \min_{\Delta U} \frac{1}{2} \Delta U^T H \Delta U + f^T \Delta U \\ \text{subject to } (M \Delta U \leq N) \end{aligned} \quad (16)$$

where:

$$\begin{aligned} H &= \Gamma^T \bar{Q} \Gamma + \bar{H}, f = (\Gamma^T \bar{Q} F) x^a - (\Gamma^T \bar{Q}) Y^*, \\ \Delta U &= [\Delta u(k) \ \Delta u(k+1) \ \dots \ \Delta u(k+N_c-1)]^T, \\ N_c &\text{ is the control horizon, } N_p \text{ is the prediction horizon,} \\ Y^* &\text{ is column vector with } p \cdot N_p \text{ elements of set points,} \\ \bar{Q} &\text{ is positive definite error weight matrix, } \bar{H} \text{ is a} \\ &\text{ } (m \times N_c) \times (m \times N_c) \text{ diagonal weighting matrix used as} \\ &\text{ tuning parameter for closed loop performance, } m \text{ is the} \\ &\text{ number of inputs,} \end{aligned}$$

$$\begin{aligned} \Gamma &= \begin{bmatrix} C_a (A_a)^0 B_a & \dots & o_{p \times m} \\ \vdots & \vdots & \vdots \\ C_a (A_a)^{N_p-1} B_a & \dots & C_a (A_a)^{N_p-N_c} B_a \end{bmatrix}, \\ F &= \begin{bmatrix} (C_a A_a)^T & \dots & (C_a (A_a)^{N_p})^T \end{bmatrix}^T, \\ M &= \begin{bmatrix} -\Phi_1 \\ \Phi_1 \end{bmatrix}, N = \begin{bmatrix} -\Delta U_{\min} + \Phi_2 \tilde{u}(k-1) \\ \Delta U_{\max} - \Phi_2 \tilde{u}(k-1) \end{bmatrix}, \Phi_1 \text{ is a} \\ &\text{ } (m \cdot N_c) \times (m \cdot N_c) \text{ lower triangular identity matrix} \end{aligned}$$

Φ_2 is a column matrix with N_c identity matrix $I_{m \times m}$, ΔU_{\min}^j and ΔU_{\max}^j , $j=1,2$ are column vectors with $(m \cdot N_c)$ elements of Δu_{\min}^j and Δu_{\max}^j .

LINEARIZING CONTROL

In this section, an exact feedback linearizing technique was used to express the forms of the commands for the considered closed loop system.

In both control scenarios it is assumed that the model is completely known, and all the state variables are available for on-line measurements. Then, taking into account that hypothesis, the following commands vanishing the dynamics of the considered outputs:

$$u_1(t) = c_1 e^{c_2 x(t)},$$

$$\begin{aligned} u_2(t) &= -(K_L a)_{CO_2} V_1 \left(\frac{x_2(t)}{K_1 10^{pH} + K_1 K_2 10^{2pH} + 1} - \frac{P x_4(t)}{H_{CO_2}} \right) - \\ &\quad - \frac{G_{in}^{N_2} x_4(t)}{x_4(t) + x_5(t) - 1}. \end{aligned}$$

Moreover, the solution of tracking problem can be solved by using a corrective term:

$$u_a(t) = [u_a^1(t) \ u_a^2(t)]^T = u(t) - \Lambda e(t), \quad (17)$$

where:

$$\begin{aligned} u(t) &= [u_1(t) \ u_2(t)]^T, \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \\ e(t) &= \begin{bmatrix} \underbrace{y_1(t) - y_1^*}_{e_1(t)} & \underbrace{y_2(t) - y_2^*}_{e_2(t)} \end{bmatrix}^T, \lambda_{1,2} > 0. \end{aligned}$$

The feedback linearizing command law (17) was chosen such that the dynamics of the tracking error $\dot{e}(t) = -\Lambda e(t)$ guarantee the exponential convergence of the outputs to the chosen setpoints y_1^* and y_2^* , respectively.

The above control law will be used in order to compare the behaviour of the closed loop system with that obtained by using the optimal command developed in the precedent section.

In accordance with the optimal command, it is considered that: $u_a(t)|_{t=kT_e} = u_a(kT_e) = u_a(k)$. Also, the applied commands are bounded, such as $u_{\min}^j \leq u_a^j(k) \leq u_{\max}^j$, $j=1,2$.

RESULTS

Suppose a photobioreactor growth process of form (12) with the following parameters (G.A. Ifrim et al. 2013; S. Tebbani et al. 2013):

$$\begin{aligned} E_a &= 172 \text{ m}^2 \text{ kg}^{-1}, E_s = 870 \text{ m}^2 \text{ Kg}^{-1}, b = 0.0008 (-), \\ K_I &= 120 \text{ } \mu\text{mol m}^2 \text{ s}^{-1}, \mu_{\max} = 0.16 \text{ h}^{-1}, \\ L &= 0.04 \text{ m}, \mu_s = 0.013 \text{ h}^{-1}, q_0 = 300 \text{ } \mu\text{mol m}^{-2} \text{ s}^{-1}, \\ M_x &= 27.8 \text{ gC - mole}, Q_p = 1.107, (K_L a)_{O_2} = 0.9 \text{ h}^{-1}, \\ (K_L a)_{CO_2} &= 0.72 \text{ h}^{-1}, K_1 = 10^{-6.35} (-), K_2 = 10^{-10.3} (-), \\ T_a &= 298.15 \text{ K}, V_I = 1.47 \cdot 10^{-3} \text{ m}^3, pH = 7.5, \\ c_3 &= 0.28 (-), V_g = 0.1764 \cdot 10^{-3} \text{ m}^3, P = 1.1013 \cdot 10^5 \text{ Pa}, \\ R &= 8.3145 \text{ J mol}^{-1} \text{ K}^{-1}, H_{O_2} = 8.384 \cdot 10^4 \text{ Pa m}^3 \text{ mol}^{-1}, \\ H_{CO_2} &= 2903.8 \text{ Pa m}^3 \text{ mol}^{-1}, G_{in}^{N_2} = 10 \text{ mL min}^{-1}, \\ C_{TIC,i} &= 0.02 \text{ mol L}^{-1}. \end{aligned}$$

The approximation (10), was obtained for the above numerical values, by using the *fit* Matlab function, so that: $c_1 = 0.1141 (-)$ and $c_2 = -1.37 (-)$.

For the predictive control algorithm, it was defined a discrete model, starting from the equilibrium values corresponding to optimal dilution (G.A. Ifrim et al. 2013):

$$\begin{aligned} \bar{x}_1 &= 0.4 \text{ g L}^{-1}, \quad \bar{x}_2 = 0.017 \text{ mol L}^{-1}, \\ \bar{x}_3 &= 9.545 \cdot 10^{-4} \text{ mol L}^{-1}, \quad \bar{x}_4 = 0.038 (-), \\ \bar{x}_5 &= 0.047 (-), \quad \bar{u}_1 = 0.05 \text{ h}^{-1}, \quad \bar{u}_2 = 1.97 \cdot 10^{-3} \text{ mol h}^{-1}. \end{aligned}$$

Also, $T_e = 0.3 \text{ h}, N_p = 8, N_c = 5,$

$$\begin{aligned} \bar{Q} &= \begin{bmatrix} 100 \cdot I_{N_p \times m N_p} & 0.001 \cdot I_{N_p \times m N_p} \end{bmatrix}^T, \\ \bar{H} &= 10 \cdot I_{m N_c \times m N_c}, \quad \Delta u_{\min}^1 = 0, \quad \Delta u_{\max}^1 = 0.1 \text{ h}^{-1}, \\ \Delta u_{\min}^2 &= 0, \quad \Delta u_{\max}^2 = 3 \cdot 10^{-2} \text{ mol h}^{-1}. \end{aligned}$$

The gains used in relation (17) are defined as: $\lambda_1 = 1.5,$
 $\lambda_2 = 0.01.$ Also, $u_{\min}^j = \Delta u_{\min}^j, u_{\max}^j = \Delta u_{\max}^j, j = 1, 2.$

All control scenarios were performed with the following initial conditions (G.A. Ifrim et al. 2013):
 $x_1(0) = 0.36, x_2(0) = 0.02, x_3(0) = 0, x_4(0) = 0.005,$

$$x_5(0) = 0.005.$$

The control schemes were implemented in Matlab/Simulink environment (Simulink 2015). The minimization problem (15) was solved by using the *quadprog* Matlab function (T. Coleman and Y. Zhang 2014).

The obtained results are presented in Figures 1, 2, 3, 4 and 5. In Figures 1 and 2, it is shown the evolutions of the considered outputs – the concentration of the biomass and the molar fraction of CO_2 for the imposed references.

Both control algorithms ensure that the tracking error converge to zero. In what concerns the biomass concentration identical results were obtained in both control scenarios. If the setpoint has an ascending profile

the biomass concentration reaches its setpoint around 12h and this time increase with descending profile of setpoint. In case of the second considered output, it was observed an overshoot for predictive control algorithm.

In Figures 3 and 4 are presented the time evolution of the control inputs applied to the system (optimal and linearizing commands).

The time evolution of the dissolved oxygen concentration is represented in Figure 5.

All the presented tests were carried out considering a constant incident light q_0 .

CONCLUSIONS

The paper presents two control strategies for a photobioreactor growth process – predictive and feedback linearizing. Starting from the highly nonlinear model of the process, it was defined an appropriate approximation in order to reduce the inhereed complexity. For predictive control strategies we define based on Jacobian and Taylor expansion technique a linear model valid around chosen equilibrium points. The control input are computed by minimization of an quadratic problem with respect to the dynamic of the process and the constraint on input. In the second control scenario an exact feedback linearizing commands are defined, based on the nonlinear model of the process.

The results from numerical simulation have prove the effectiveness of the proposed algorithms: the output variables follows the imposed references.

The slightly better results were obtained in the second control scenario. However, the main drawback of the method can be the constant matrix gain Λ , which is not suitable for the design of a robust controller (for example the incident light may varies according with the environment conditions).

Instead, the proposed predictive controller was proven their robustness against the external disturbances of the model (see F. Stinga and all, 2015)

So, the future work shell concerns to prove the robustness analyses of the presented control strategies against the uncertainites of the model (the variation of the incident light) and, also for practical implementation of the proposed control schemes.

ACKNOWLEDGMENT

This work was supported by UEFISCDI, project BIOCON no. PN-II-PT-PCCA-2013-4-0070, 269/2014.

REFERENCES

- F. Angulo, G. Olivar, and A. Rincon, "Control of an anaerobic upflow fixed bed bioreactor," in *Proc. 15th Mediterranean Conf. on Contr. & Autom.*, July 27-29, 2007, Athens, Greece, Paper T04-005.
- G. Bastin and D. Dochain, *On-line estimation and adaptive control of bioreactors*, New York, NY: Elsevier, 1990.
- O. Bernard (Responsible), *Design of models for abnormal working conditions and destabilisation risk analysis*. Report Number: D3.1b, TELEMAC IST 2000-28156, 2004, 81 pages.
- O. Bernard, "Hurdles and challenges for modelling and control of microalgae for CO2 mitigation and biofuel production," *J. Process Control*, vol. 21, pp. 1378–1389, 2011.

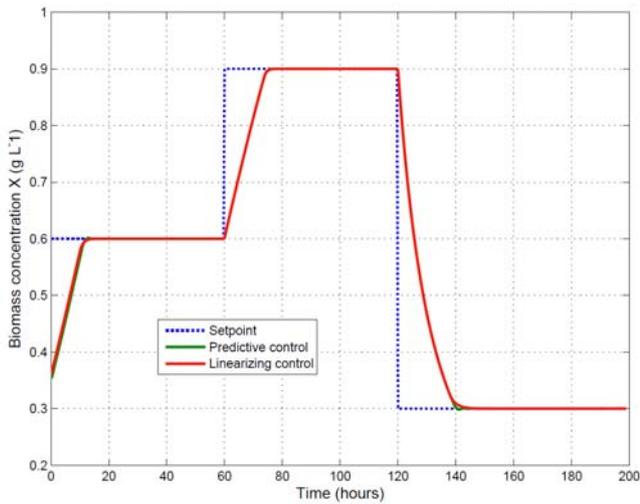


Figure 1: The time evolution of the biomass concentration

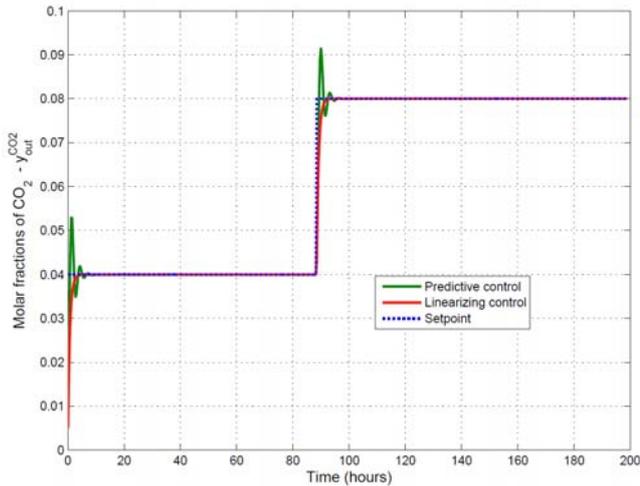


Figure 2: The time evolution of the molar fraction of CO_2

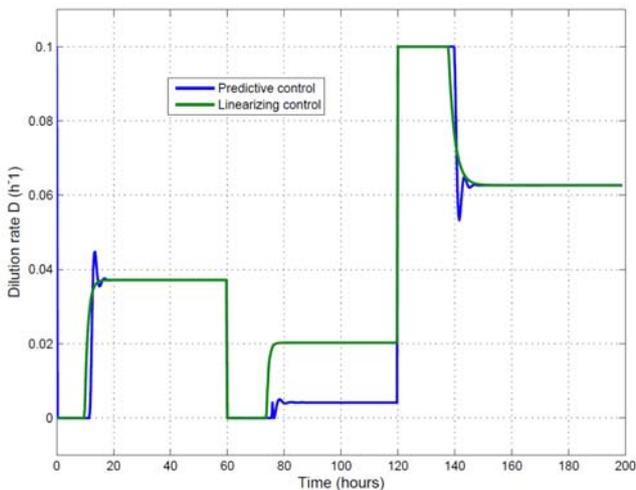


Figure 3: Profile of the command input D

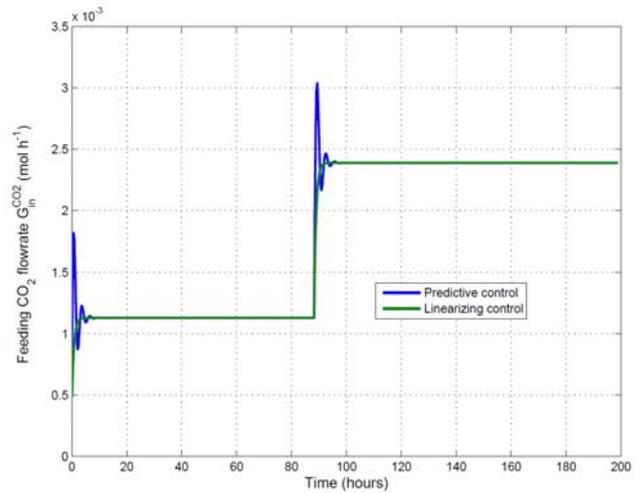


Figure 4: The time evolution of the command input $G_{in}^{CO_2}$

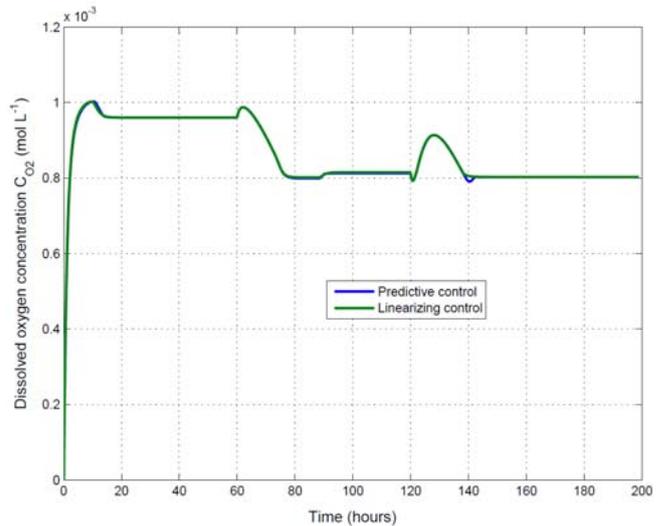


Figure 5: The time evolution of the dissolved oxygen concentrations C_{O_2}

G.A. Ifrim, Control of Two Biological Processes of Environmental Interest (Biological Wastewater Treatment and Microalgae Production in Photobioreactor), PhD thesis, University of Nantes, France or University "Dunarea de Jos", Galati, Romania, 2012.

G.A. Ifrim et al., "Multivariable feedback linearizing control of Chlamydomonas reinhardtii photoautotrophic growth process in a torus photobioreactor," *Chemical Eng. Journal*, vol. 218, pp. 191–203, 2013.

S. Tebbani, M. Titica, S. Caraman, and L. Boillereaux, "Estimation of Chlamydomonas reinhardtii Growth in a Torus Photobioreactor," Preprints of the 12th IFAC Symposium on Computer Applications in Biotechnology, 16-18 Dec., 2013, Mumbai, India, pp. 155–160.

S. Tebbani, F. Lopes, R. Filali, D. Dumur, and D. Pareau, "Nonlinear predictive control for maximization of CO_2 bio-fixation by microalgae in a photobioreactor," *Bioprocess Biosyst. Eng.*, vol. 37, no. 12, pp. 83–97, 2014.

S. Tebbani, F. Lopes, and G. BecerraCelis, "Nonlinear control of continuous cultures of *Porphyridium purpureum* in a photobioreactor," *Chemical Engineering Science*, vol. 123, no. 18, pp. 207–219, 2015.

D. Dochain and P. Vanrolleghem, *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. IWA Publ., 2001.

D.J. Batstone et. al, The IWA Anaerobic Digestion Model No 1 (ADM1), *Water Sci. & Technology*, vol. 45, no. 10, pp. 65-73, 2002.

F. Mairet, O. Bernard, M. Ras, L. Lardon, and J.P. Steyeret, "Modeling anaerobic digestion of microalgae using ADM1," *Bioresource Technology*, vol. 102, pp 6823–6829, 2011.

F. Mairet, O. Bernard, M. Ras, L. Lardon, and J.P. Steyeret, "A Dynamic Model for Anaerobic Digestion of Microalgae," Preprints of the *18th IFAC World Congress, Milano (Italy)*, Aug. 28 – Sept. 2, 2011, pp 5034–5039.

D. Dochain, Ed., *Automatic Control of Bioprocesses*. UK: ISTE / John Wiley & Sons, 2008.

E. Petre, D. Selişteanu, and D. Şendrescu, "Adaptive and robust-adaptive control strategies for anaerobic wastewater treatment bioprocesses," *Chem. Eng. J.*, vol. 217, pp. 363-378, 2013.

I. Neria-González, A.R. Dominguez-Bocanegra, J. Torres, R. Maya-Yescas, and R. Aguilar-López, "Linearizing control based on adaptive observer for anaerobic continuous sulphate reducing bioreactors with unknown kinetics," *Chem. Biochem. Eng. Q.*, vol. 23, no. 2, pp. 179–185, 2009.

F. Logist, B. Houska, M. Diehl, and J.F. Van Impe, "Robust multi-objective optimal control of uncertain (bio)chemical processes," *Chem. Eng. Sci.*, vol. 66, no. 20, pp. 4670–4682, 2011.

D. Selişteanu, E. Petre, and V. Răsvan, "Sliding mode and adaptive sliding mode control of a class of nonlinear bioprocesses," *Int. J. Adapt. Contr. & Signal Process*, vol. 21, no. 8-9, 2007, pp. 795-822.

M. Morari and J.H. Lee, "Model predictive control: past, present and future", in *Computers and Chemical Engineering*, 23(4-5), pp. 667-682, 1999.

Simulink User's Guide, http://www.mathworks.com/help/pdf_doc/simulink/sl_using.pdf, 2015.

T. Coleman and Y. Zhang, *Optimization Toolbox User's Guide*, The MathWorks Inc., available online at http://www.mathworks.com/help/pdf_doc/optim/optim_tb.pdf, 2015.

F. Stinga, M. Marian and A. Soimu, "Online estimation and control of an induction motor", in *Proc. Of the 19th International Conference on System Theory, Control and Computing*, pp. 742-746, 2015.



FLORIN STÎNGĂ 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 Lecturer in the Department of Automatic Control at the Faculty of Automation, Computers and Electronics, Craiova. His researches interested are in hybrid dynamical systems and predictive control.



EMIL PETRE received the Engineering degree in Automatic Control and Computers in 1977 and Ph.D. degree in Automatic Control in 1997 from University of Craiova, Romania. Since 1981 he is with the University of Craiova. Currently, he is Professor at the Department of Automatic Control, and from 2011, Dr. Petre serves as director of this department. His research interests include nonlinear systems, adaptive control, estimation and identification, control of bioprocesses, and neural control.

GREENHOUSE MODELING AND SIMULATION FRAMEWORK FOR EXTRACTING OPTIMAL CONTROL PARAMETERS

Byeong Soo Kim, Bong Gu Kang, and Tag Gon Kim
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, 305-701, Republic of Korea
E-mail: {bskim, bgkang}@smslab.kaist.ac.kr
tkim@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

Greenhouse control, system identification, neural networks, optimization.

ABSTRACT

In a greenhouse system, a control is important to allow optimal growth conditions for crops. However, because testing the greenhouse for real conditions requires much time and money, the modeling-and-simulation approach is necessary to predict and improve the greenhouse environment. There is much research related to greenhouse control, there is a lack of research on applicable frameworks for real greenhouses. Therefore, this paper proposes a greenhouse modeling-and-simulation framework to extract optimal control parameters. The proposed work is composed of three parts: system identification, controller design, and optimization. The plant model is built through system identification, and the model is controlled by the controller, which is affected by disturbances. This simulation is repeated through design of experiments to optimize the control parameters. This paper presents an experiment with real greenhouse data from Jinju, Korea to show the usefulness of the proposed framework. It gives insight into the decision of choosing control parameters and helps to raise agricultural productivity.

INTRODUCTION

Greenhouse control to create a favorable environment to improve crop development is an important problem. Proper control can help maximize the productivity of crops. Thus, maintenance of environmental parameters, like greenhouse indoor temperature, humidity, CO₂ levels, and so on, according to the plant growth cycle, is required. There are many elements used to control these parameters in the greenhouse control system, for example ventilators (or windows), heaters, or shading screens. However, it is difficult to find optimal control parameters according to the specifications of greenhouses because constructing real greenhouses requires much time and can be expensive.

Modeling and simulation (M&S) can be the best method to overcome this problem. We can easily extract the optimal parameter set for a greenhouse through the M&S framework. Then, we can obtain the optimal growth conditions for crops by applying the parameter set into

the real control system (Figure 1). Because the greenhouse control system is too complex to model completely, it is necessary to analyze and classify the system according to the objective of M&S. The greenhouse system is largely composed of plant, sensor, controller, actuator, and environment. As depicted in Figure 1, they work as follows. The operation result occurs through the controller and actuator when control parameters are set up in the controller. Then, the plant model generates outputs (indoor temperature and humidity). The outputs reach the controller as a feedback. Such a process is executed in every clock recursively, and consequently, the plant outputs can be adjusted close to the set point that we want to acquire.

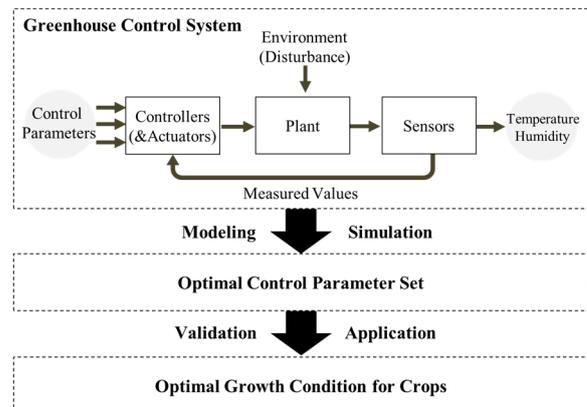


Figure 1: Overview of Greenhouse Control System

There is a glut of research related to the modeling of greenhouse control systems. The details will be described in the next section. However, there is a lack of research providing an overall framework that is possible to apply directly to a real greenhouse. Therefore, this paper proposes a greenhouse M&S framework to extract optimal control parameters. The control parameters are important elements that decide the specification and performance of the controller. The proposed framework includes the entire process, including system identification, along with control and parameter optimization.

This paper is organized as follows. Previous works related to our study are briefly introduced. Then, our proposed framework for extracting optimal control parameters using a neural network is described. Finally, an experiment with real greenhouse data is provided.

RELATED WORK

There has been much research regarding modeling greenhouse control systems. Some research has focused on how to model and predict the greenhouse model (Cunha 2003). This research applied several methods, like physical modeling (Bot 1991), autoregressive exogenous (ARX) modeling, and artificial neural network (ANN), to create plant models. They showed pros and cons of each modeling approach, but they did not consider the part of the controller model and the parameter optimization of the greenhouse environment.

In control fields, there has been research that applied several techniques to control the greenhouse climate (Hagan and Demuth 1999). A proportional-integral-derivative (PID) controller is one of the representative methods to control the feedback system, and it is frequently adapted to the greenhouse model (Cunha et al. 1997). Other researchers have applied robust adaptive control to implement the controller (Bennis et al. 2008; Luan et al. 2011). Robust control is an approach that refers to the control of uncertain plants with unknown dynamics subject to uncertain disturbances (Chandraseken 1996). They build the plant model through physical modeling, then implement a controller through the theory of robust control.

These researches provides controllers with high performance. However, they do not consider the disturbances that vary with time. Also, they are hard to be implemented and applied to the real greenhouse control system directly. There is no research on how to provide a practical framework that is applicable to a real greenhouse system. For this, the framework should include all of plant modeling, controller modeling, and parameter optimization.

Therefore, we propose an M&S framework for extracting greenhouse control parameters that is applicable to the real greenhouse system in this paper. The next section will describe this proposed framework in detail.

GREENHOUSE M&S FRAMEWORK

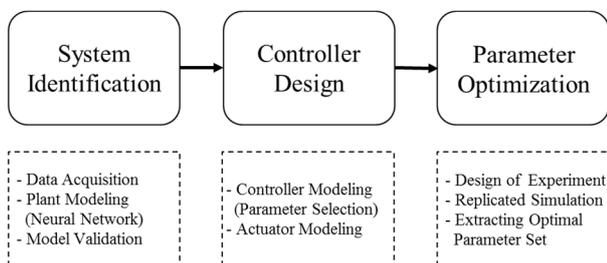


Figure 2: Overall Process of Proposed Framework

In this section, we propose a greenhouse M&S framework for extracting optimal control parameters. This proposed framework helps to draw optimal parameters by acquiring greenhouse data, regardless of the specification of the greenhouse. As shown in Figure 2, it is mainly composed of three parts: system identification, controller design, and parameter

optimization. In the system identification step, we built a plant model with acquired greenhouse data using ANN. In the control step, the temperature and humidity were controlled by the control algorithm. Finally, in the parameter optimization step, we drew the optimal control parameter set through repeated simulations designed by design of experiments (DOE).

In this paper, our target system is a greenhouse located in Jinju, Korea. Its specification is concretely depicted in Table 1. Our proposed framework is applicable to this greenhouse as mentioned below.

Table 1: Specification of Greenhouse

Type	Specification
Location	Jinju, South Korea
Dimensions	30 x 90 x 7.5 (m)
Date / sampling time	2015, April~June / 1 minute
Kind of crop	Tomato
Sensors	Temperature, humidity, CO ₂ , wind speed, light density, etc.
Actuators	Window, heater, screen, etc.

Part 1: System Identification

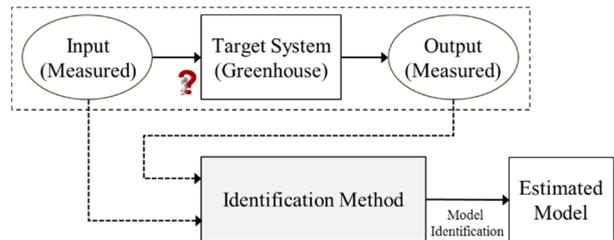


Figure 3: Basic Concept of System Identification

System identification is a process used to build a mathematical model of the dynamics of a system from measured data (Nelles 2000). Figure 3 shows the concept of system identification. The design of the control system requires a system-identification process to build a model of the dynamics.

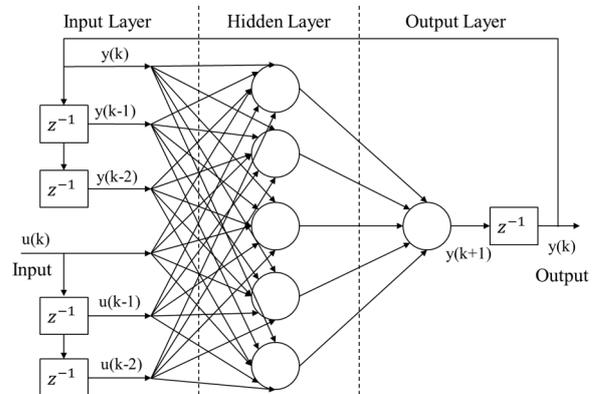


Figure 4: Tapped-Delay-Line Neural Network

There are several methods of system identification (Ljung 1987), like ANN. We used ANN to identify the plant model of the greenhouse in the proposed work (Narendra and Parthasarathy 1990). ANN is a machine learning approach that is inspired by biological neural networks and composed of a large number of highly interconnected neurons. During the learning, the strengths of neuron connection (weights) are changed in order to calibrate the model. ANN can predict the future behavior of the system precisely (Sjöberg et al. 1994). As shown in Figure 4, we use tapped-delay-line neural network (TDLNN) for the prediction of time series (Gupta et al. 2004).

In our system identification, we first acquired and analyzed the observational greenhouse data to use them as training data. They were classified into inputs, outputs, and disturbances. It was important to include parameters mainly affecting the plant model and exclude the other minor ones before we identified the plant model. In this paper, we use greenhouse parameters as depicted in Table 2 to identify the model through TDLNN. Control inputs and disturbances were used for the inputs of neural network, and control outputs were used for the outputs of neural networks.

Table 2: Parameters Used in Plant Model

Type	Parameter	Description
Control Inputs	Pw (%)	Window Angle
	Ht (%)	Heater
Control Outputs	Ti (°C)	Indoor Temperature
	Hi (°C)	Indoor Humidity
Disturbances	To (°C)	Outdoor Temperature
	Ql (W/m ²)	Light Quantity
	Sw (m/s)	Wind Speed

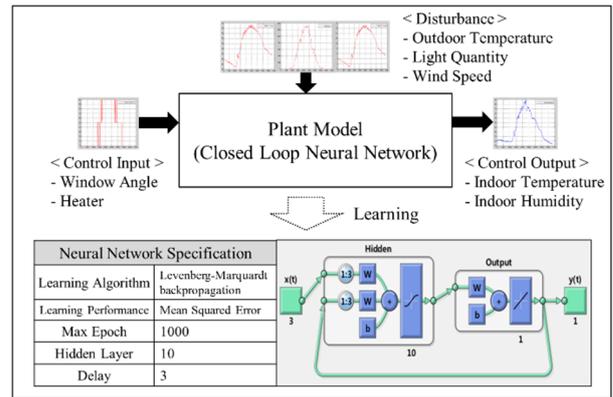


Figure 5: System Identification Process

In this paper, we use the Lavenberg-Marquardt optimization technique as a learning algorithm (Marquardt 1963), and mean squared error for a measurement of learning performance. Figure 5 shows the identification of the plant model and its specification. Also, Figure 6 represents the comparison result between real indoor temperature and predicted indoor temperature using the identified plant model. It shows that the identified plant model reflects well the real greenhouse plant.

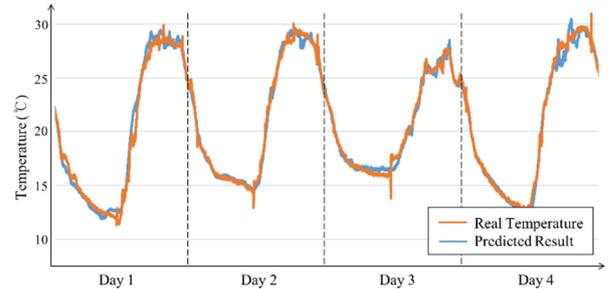


Figure 6: Result of System Identification

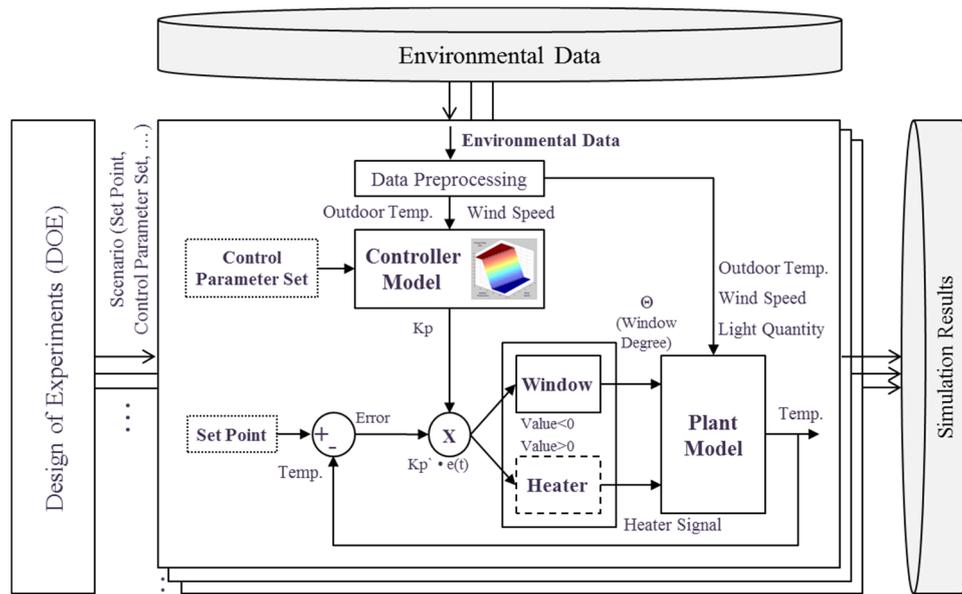


Figure 7: Overall M&S Framework

Part 2: Controller Design

After the process of system identification, we used the plant model to control the greenhouse system. Even though there were many elements to be controlled, we focused on the window and heater. They played an important role in decreasing and increasing inner temperature and humidity. In this paper, we only considered the window control in order to focus on one result. A P controller is generally used to control the window in the greenhouse. The proposed framework applied the concept of a P-band for the P control (Kamp 1996). A P-band determines the opening angle of the window according to the difference between the measured point and the set point. It is the reciprocal of the proportional gain constant (K_p) generally used in the P controller.

Table 3: Description of Control Parameter Set

Parameter	Description
max.out.temp	Maximum threshold of outdoor temp.
min.out.temp	Minimum threshold of outdoor temp.
max.wind.spd	Maximum threshold of wind speed
min.wind.spd	Minimum threshold of wind speed
max.pband	Maximum threshold of P-band
min.pband	Minimum threshold of P-band
deg.win.open	Window opening angle with one execution

In this paper, we used a modified P controller that had a variable P-band ($= 100 / K_p$) value, not a fixed P-band. The controller reflected outdoor temperature and wind speed from the greenhouse data dynamically because the P-band value is influenced by the parameters. (Kamp 1996) For example, the P-band value decreases when the temperature rises, and conversely, it increases when the wind speed increases. The variable P-band graph determined by outdoor temperature and wind speed is shown in Figure 8. However, it is hard to know the optimal shape of the graph (including gradient, maximum, and minimum threshold value) that can maximize the growth of the crops. So, the optimization step is required to find the optimal control parameters that determine the shape of the graph.

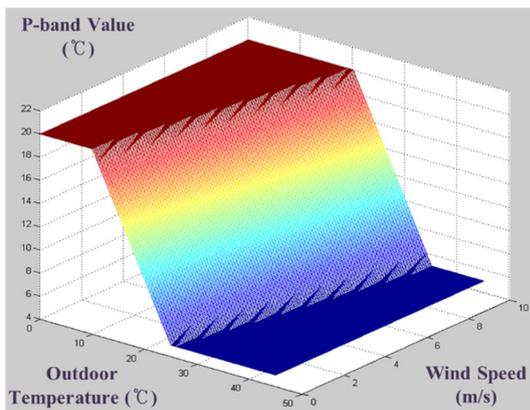


Figure 8: Decision of P-band Value

Part 3: Parameter Optimization

After design and implementation of the controller, parameter optimization is required to draw an optimal control parameter set. The control parameters are the seven input parameters that determine the shape of the P-band graph (Figure 8), as shown in Table 3. They are actual parameters used in the greenhouse control system, using the controller as mentioned previously. When the shape of the graph changes due to these input parameters, indoor temperature and humidity are also affected. Therefore, replicated simulations should be performed in accordance with the designs that are made by DOE. There are various methods of DOE, like full-factorial design, central composite design, and Box-Behnken design, used to perform efficient experiments (Antony 2003). Using these DOE methods, we can find an optimal control parameter set (input), which has the lowest root mean squared error (RMSE) value (output). Figure 7 represents the overall M&S framework, including all of the parts.

EXPERIMENTS

In this section, we represent the experiments using the real greenhouse data to show the effect of the proposed work. We acquired the data from the greenhouse located in Jinju as mentioned earlier (Table 1).

Experimental Design

We designed experiments to acquire the optimal control parameter set, which would have the lowest RMSE value. To find the optimal control parameter set, we simply used full-factorial design. Table 4 shows seven parameters and their values in the greenhouse control system. A total of 15,625 simulation runs were required to find the optimal parameter set.

We used MATLAB/Simulink to implement our proposed work. We also used the six thousand sample data over four days to simulate the greenhouse control system, as depicted in Table 1.

Table 4: Parameter Set of Greenhouse Control System

Parameter (Input)	Value	Number
max.out.temp (°C)	23 ~ 27	5
min.out.temp (°C)	16 ~ 21	5
max.wind.spd (m/s)	1 ~ 5	5
min.wind.spd (m/s)	0	1
max.pband (°C)	18 ~ 22	5
min.pband (°C)	3 ~ 7	5
deg.win.open (%)	10, 15, 20, 25, 30	5
Total Executions	15,625 ($=5^6$)	

Experimental Results

We obtained an optimal control parameter set, which had the lowest RMSE value, through replicated simulations with DOE (Table 5). The RMSE of the simulation with the optimal set was 2.511, and the RMSE of real data was

4.195. That is, the adaptation of the optimal set led to a 40.1% performance improvement compared with real temperature, which did not need to be optimized. Figure 9 shows the graph of the control result using this parameter set. In this experiment, the performance improvement is only showed in the part of cooling, which is controlled by the window, because we did not use the control data of the heater.

Table 5: Simulation Result: Optimal Parameter Set

Parameter	Optimal Value
max.out.temp (°C)	25
min.out.temp (°C)	19
max.wind.spd (m/s)	1
min.wind.spd (m/s)	0
max.pband (°C)	19
min.pband (°C)	4
deg.win.open (%)	10
RMSE improvement	40.1%

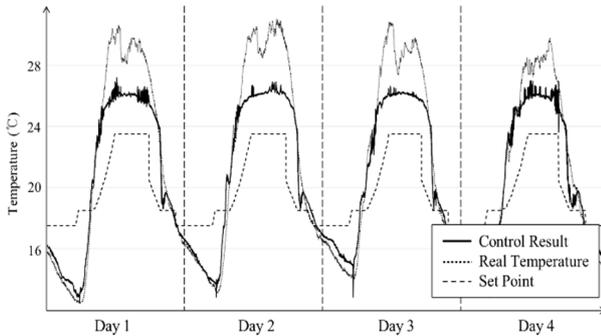


Figure 9: Simulation Result: Predicted Temperature

The result shows that the indoor temperature was controlled nearer to the set point through the adaptation of the optimal parameter set. That is, periodic simulation and adaptation of the parameter set can help to maintain more suitable indoor temperatures for the crops without any trial and error. The user only needs to input the parameter set into the greenhouse control system. Consequently, we know that the proposed M&S framework can provide optimal growth conditions by predicting the greenhouse environment.

CONCLUSIONS

This paper proposed a greenhouse M&S framework to draw optimal control parameter set. It is hard to test the real greenhouse due to the time needed and cost; thus, the M&S approach was applied to predict and improve the greenhouse environment. The proposed work was composed of three parts: system identification, controller design, and optimization. The plant model was built through system identification, and the model was controlled by the controller, which was affected by certain disturbances. This control simulation was repeated with DOE to optimize the control parameters.

In this paper, we used the data acquired from the greenhouse in Jinju, Korea to show the usefulness of the proposed framework. We drew the optimal control parameter set by applying the data taken over four days to the framework, which allowed us to verify that it had an improved control performance. This result means that the framework can give insight into the decision of control parameters and raise agricultural productivity. For further work, we will fully automate the each step of the framework to adapt it to the real greenhouse system. Also, we will apply various techniques about neural network, control, and optimization for the performance improvement.

REFERENCES

- Antony, J. 2003. *Design of Experiments for Engineers and Scientists*, Butterworth-Heinemann.
- Bennis N.; J. Duplaix; G. Enea; M. Haloua; and H. Youlal. 2008. "Greenhouse Climate Modelling and Robust Control." *Computer and Electronics in Agriculture* 61, 96-107.
- Bot, G.P.A. 1991. "Physical Modelling of Greenhouse Climate." *In: Proceedings of the IFAC/ISHS Workshop*, 7-12.
- Chandrasekharan, P. C. 1996. *Robust Control of Linear Dynamical Systems*, Academic Press.
- Cunha J.B.; C. Couto; and A.E. Ruano. 1997. "Real-time Parameter Estimation of Dynamic Temperature Models for Greenhouse Environmental Control." *Control Engineering Practice* 5, No.10, 1473-1481.
- Cunha, J.B. 2003. "Greenhouse Climate Models: An Overview." *The 3th Conference of the European Federation for Information Technology in Agriculture* (Debrecen, Hungary, Jul.5-9).
- Gupta, M.; L. Jin; and N. Homma. 2004. *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory*, John Wiley & Sons.
- Hagan, M.T. and H.B. Demuth. 1999. "Neural Networks for Control." *In Proceedings of the American Control Conference* (San Diego, CA, Jun.2-4). IEEE, 1642-1656.
- Kamp, P.G.H.; and G.J. Timmerman. 1996. *Computerized Environmental Control in Greenhouses*, IPC-Plant.
- Ljung, L. 1987. *System Identification - Theory for The User*, PTR Prentice-Hall.
- Luan X.; P. Shi; and F. Liu. 2011. "Robust Adaptive Control for Greenhouse Climate Using Neural Networks." *International Journal of Robust and Nonlinear Control* 21, 815-826.
- Marquardt, D.W. 1963. "An algorithm for least-squares estimation of nonlinear parameters." *Journal of the society for Industrial and Applied Mathematics* 11, No.2, 431-441.
- Narendra K.S. and K. Parthasarathy. 1990. "Identification and Control of Dynamical Systems Using Neural Networks." *Neural Networks, IEEE Transactions on* 1, No.1, 4-27.
- Nelles, O. 2000. *Nonlinear System Identification*, Springer.
- Sjöberg, J.; H. Hjalmarsson; and L. Ljung. 1994. *Neural Networks in System Identification*.

AUTHOR BIOGRAPHIES

BYEONG SOO KIM is a PhD 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), data modeling, and big data.

BONG GU KANG is a PhD. 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), communication simulator, and interoperation.

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 has published about 200 papers on M&S theory and practice in

international journals and conference proceedings. He is very active in defense modeling and simulation in Korea. He is a fellow of SCS and a senior member of IEEE.

HAE SANG SONG studied his MS.D and PhD courses in electrical engineering at the Korea Advanced Institute of Science and Technology (KAIST). He worked for a couple of years in an R&D lab, Institute of Advanced Engineering (IAE), in 1999–2000. He also worked at a venture company for about two years and has been a professor of Dept. Computer Engineering at Seowon University, Korea, since 2002. His major interests reside in modeling simulation, analysis, and control of discrete-event dynamic systems.

DOMINANT SPECTRUM ASSIGNMENT FOR NEUTRAL TIME DELAY SYSTEMS: A STUDY CASE

Libor Pekař and Roman Prokop
Faculty of Applied Informatics, Department of Automation and Control Engineering
Nad Stráněmi 4511, 76005 Zlín, Czech Republic
E-mail: {pekar, prokop}@fai.utb.cz

KEYWORDS

Characteristic Quasipolynomial, Finite Spectrum Assignment, MATLAB[®]/Simulink[®], Neutral System, Time Delay System.

ABSTRACT

This paper is aimed at the presentation of a suboptimal numerical algorithm that has been designed for the assignment of a finite dominant part of the infinite eigenvalue spectrum of a strongly stable neutral time delay control system. Once a structure of a conventional finite-dimensional controller for a delayed controlled plant is suggested, desired loci of a sufficient number of feedback poles are selected, according to required dynamical properties. The goal is to bring the dominant spectrum to prescribed one as close as possible. The first step of the procedure insists in the use of the Quasi-Continuous Shifting Algorithm (QCSA). Then, the results are enhanced by an optimization procedure for a suitable objective function. The presented methodology is demonstrated by a simulation example of control of an unstable time delay system (TDS) in the MATLAB[®]/Simulink[®] environment. Some possibilities how to adjust the algorithm are given to the reader as well.

INTRODUCTION

Although there have been derived and designed various unconventional ad-hoc control methods and strategies for linear time-invariant time delay systems (TDS) in the scientific literature during the early years of this millennium, see e.g. (Chiasson and Loiseau 2012; Richard 2003; Sipahi et al. 2012), the use of conventional proportional-integral-derivative (PID) controllers still plays a crucial role in modern control theory despite the made progress and advances (Åström and Hägglund 2006; Wang et al. 2009; Zitek et al. 2013).

If, however, a finite-dimensional controller is applied to a TDS plant, an infinite-dimensional control feedback system is obtained. This feature is characterized by the fact that the eventual characteristic quasipolynomial instead of a polynomial (the zeros of which usually agree with system poles) includes exponential terms. Simply, a PID law can not cancel delays in the feedback loop. In such cases, the task of controller tuning yields the problem of a suitable setting of a finite number of

adjustable controller parameters faced with the infinite spectrum of system poles. There is a natural effort to develop tuning procedures for the aforementioned class of systems which are usable and understandable also for non-experts without an excessive mathematical formulation. A possible way is to shape the dominant the feedback spectrum by means of pole placement controller parameters tuning principles.

A one-shot or direct pole assignment for controllable TDS has been presented e.g. in (Lee and Zak 1982; Zitek and Vyhliđal 2002). A more advanced idea is based on successive shifting the dominant poles to the left (stable) complex half-plane by using the Quasi-Continuous Shifting Algorithm (QCSA) (Michiels et al. 2002; Michiels and Vyhliđal 2005), or other methods (Michiels and Gumussoy 2014; Vyhliđal 2003). However, all these methods intend to minimize the spectral abscissa only. The pole-matching problem for retarded TDS operating in the state-space has been solved in (Michiels et al. 2010) where poles can not leave the prescribed positions and the unrestrained rest of the spectrum is attempted to be pushed to the left, which may results in a lengthy trial-and-reset placing procedure.

In (Pekař and Navrátil 2014), we introduced an algorithm called the PPSA (Pole-Placement Shifting based controller tuning Algorithm) for retarded TDS where both poles and zeros are selected according to desired closed-loop dynamic properties represented by the finite-dimensional model. During the shifting procedure minimizing both spectral abscissas as a secondary objective function, poles and zeros can leave their prescribed positions but remain in their vicinities.

In this paper, ideas and methodology of the PPSA are applied to dominant low-frequency pole assignment in input-output neutral TDS model formulation. Neutral TDS spectral properties are more advanced, tricky and intricate compared to retarded ones, i.a. the so-called strongly stable system are to be reached. First, dominant poles are forced to move towards the prescribed positions. Then, the objective function reflecting the distance of prescribed poles from the actual ones and the abscissa of the rest of the spectrum is minimized by means of the Nelder-Mead technique (Nelder and Mead 1965). The whole procedure is simply implementable in standard program languages.

A detailed simulation example performed in the MATLAB[®]/Simulink[®] environment provides the reader with the procedure demonstration and performance verification.

PRELIMINARIES

Neutral TDS Spectral Properties

Let basic spectral and exponential stability properties of retarded and neutral TDS be introduced first. Consider a single-input single-output (SISO) TDS governed by the following transfer function

$$G(s) = N(s)/D(s) \quad (1)$$

where $N(s)$, $D(s)$ are quasipolynomials of the general form $X(s) = s^n + \sum_{i=0}^n \sum_{j=1}^{h_i} x_{ij} s^i e^{-s \sum_{k=1}^L \lambda_{ij,k} \tau_k}$ in which $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_L) \in \mathbb{R}_+^L$, represent independent delays, $\lambda_{ij,k} \in \mathbb{N}_0$ and $x_{ij} \in \mathbb{R}$. Let $X_a(s) = 1 + \sum_{j=1}^{h_n} x_{nj} e^{-s \sum_{k=1}^L \lambda_{nj,k} \tau_k}$ be the associated exponential polynomial related to $X(s)$. If $D_a(s) \in \mathbb{R}$, system (1) is called as retarded; otherwise, the system is of a neutral type.

Assumption 1. Assume that

$$\Sigma_{ND} = \{s_{ND,k}\} = \emptyset, s_{ND,k} := \{s : N(s) = D(s) = 0\}$$

that is, there are no common roots of $N(s)$, $D(s)$.

Under Assumption 1, the roots of $D(s)$ coincide with system poles. For their spectrum, Σ , it holds the following properties (Hale and Verduyn Lunel 1993; Michiels and Niculescu 2007).

Property 1. For system (1) of neutral type it holds that:

1. If there exists a nonzero pair $\{a_{ij}, \lambda_{ij,k}\}$ for some $\tau_k > 0$ and some i, j, k , then $|\Sigma| = \infty$.
2. There exists a vertical chain of poles, s_k , at $\gamma := \sup \operatorname{Re} \Sigma_a$ such that $\lim_{k \rightarrow \infty} \operatorname{Re} s_k = \gamma$, $\lim_{k \rightarrow \infty} \operatorname{Im} s_k = \infty$ where Σ_a is the zero set of $X_a(s)$.
3. Isolated poles behave continuously and smoothly with respect to $\boldsymbol{\tau}$ on \mathbb{C} .

However, rather different features hold for the so-called spectral abscissa defined as $\alpha(\mathbf{p}) := \mathbf{p} \mapsto \sup \operatorname{Re} \Sigma$, for a parameters vector \mathbf{p} .

Property 2. For the spectral abscissa of neutral system (1) holds the following (Vanbiervliet et al. 2008):

1. It may be nonsmooth and hence not differentiable, e.g. in points with more than one real pole or conjugate pairs with the same maximum real part.
2. It is non-Lipschitz, for instance, at points where the maximum real part has multiplicity greater than one.

Neutral TDS Stability

Among many approaches to stability of neutral TDS, exponential and strong ones are matters of this contribution.

Proposition 1 (Michiels and Vyhlídal 2005). Neutral system (1) is exponentially stable if $\alpha(\mathbf{p}) < -\varepsilon, \varepsilon > 0$.

Whereas the notion of exponential stability is well-known, the concept of strong stability is much more unfamiliar to researchers and engineers. It expresses the ability of Σ_a to persist in the left (stable) half-plane under small delay perturbations.

Definition 1 (Hale and Verduyn Lunel 1993; Michiels and Niculescu 2007; Michiels and Vyhlídal 2005). Neutral system (1) is said to be strongly stable if

$$\bar{\gamma} = \sup \{ \gamma(\boldsymbol{\tau} + \delta\boldsymbol{\tau}) : \|\delta\boldsymbol{\tau}\| < \varepsilon \} < 0$$

for any sufficiently small $\varepsilon > 0$.

Property 3 (Michiels and Vyhlídal 2005). Number of poles with $\operatorname{Re} s_k > \bar{\gamma} + \varepsilon_1$ for a sufficiently small $\varepsilon_1 > 0$ is always finite and they are isolated.

Proposition 2 (Vyhlídal 2003). The system is strongly stable if

$$\sum_{j=1}^{h_i} |d_{nj}| < 1 \quad (2)$$

When achieving exponential stability it is desirable for practical reasons to satisfy strong stability as well.

DOMINANT SPECTRUM ASSIGNMENT ALGORITHM

Consider the closed-loop control feedback system governed by the transfer function (1) with the denominator including the number of $r > 0$ selectable parameters

$$\mathbf{K} = (K_1, K_2, \dots, K_r)^T \neq \mathbf{0} \in \mathbb{R}^r$$

hence, $D(s) = D(s, \mathbf{K})$, and let Assumption 1 hold hereinafter. The designed algorithm framework solving the pole placement matching problem for neutral TDS can be summarized as follows.

Algorithm 1.

1. It is given the feedback denominator $D(s, \mathbf{K})$ as the input.
2. If $D_a(s, \mathbf{K} \equiv \mathbf{0})$ and condition (2) does not hold, abandon the algorithm; else, select the number of $n < r$ poles and their loci according to desired feedback dynamics.
3. Place a subset of poles to prescribed positions by using a direct pole placement methodology. The initial setting \mathbf{K}_0 is obtained.
4. If the placed spectrum is the rightmost (dominant) within the selected range of frequencies and (2) holds, terminate the algorithm (go to step 6); else, move the dominant roots to the desired loci by means of a shifting algorithm, and simultaneously, push the rest of the spectrum to the left as far as possible. Denote the eventual result as \mathbf{K}_i , where i expresses the achieved number of iterations.
5. If the shifting is successful (see step 4), terminate the algorithm; otherwise, minimize the cost function $\Phi(\mathbf{K})$ reflecting the distance of dominant roots from prescribed ones and the spectral abscissa of the rest of the spectrum, by using an optimization iterative algorithm. The (sub)optimal solution \mathbf{K}_{opt} is obtained.
6. Get \mathbf{K}_0 , \mathbf{K}_i or \mathbf{K}_{opt} as the output.

Algorithm Discussion

Going into details of Algorithm 1, step 2 means that if the feedback system is strongly unstable and the associated exponential polynomial can not be affected by selectable parameters, there is no sense to shape the spectrum anymore. The number of selected desired poles should be less the number of free parameters to remain some degrees of freedom to adapt $\alpha(\mathbf{K})$.

In step 3, the reader is referred e.g. to (Vyhlídal 2003; Zitek and Vyhlídal 2002) for details. Once the initial spectrum is placed, its dominant part is checked. If it concurs with the desired loci, the placement is sufficient and there is no reason to made improvements (unless the user wants to enhance the spectral abscissa). Contrariwise, rightmost poles may be successively shifted towards the prescribed positions. Here in step 4, we have to highlight our observation: Although it has been stated in (Michiels and Vyhlídal 2005) that for a neutral TDS there is no reason to deal with poles $\text{Re } s_k < \bar{\gamma}$ (see Property 1, and Definition 1), we have observed by simulations that it is desirable to control also poles left from this vertical line with a sufficiently small modulus in some cases. The idea can simply be explained as follows. Consider the vertical strip of poles introduced in Property 1, item 2, with some unperturbed γ and $D_a(s, \mathbf{K} \neq \mathbf{0})$. If a finite number of isolated poles satisfies $\text{Re } s < \gamma$ but they are right from the bunch of poles constituting the strip, the essential part

of the system dynamics might be determined by this small low-frequency subset. Moreover, the eventual value of $\bar{\gamma}$ can be adjusted. In other words, from the dynamical point of view, it is not reasonable to deal with the rightmost high-frequency poles. Nevertheless, their position must be checked with respect to exponential stability.

Step 5 of Algorithm 1 includes the optimization procedure that may be performed via several techniques. The crucial substep consists of the formulation of the objective function that must consider up to three factors: The distance of current dominant poles from the desired ones, the spectral abscissa of the rest of the spectrum, $\alpha_r(\mathbf{K})$, and the condition (2). Thus, three subfunctions $\phi(\mathbf{K})$, $\pi(\mathbf{K})$, $\varphi(\mathbf{K})$ are to be set, respectively, such that

$$\Phi(\mathbf{K}) = \phi(\mathbf{K}) + \pi(\mathbf{K}) + \varphi(\mathbf{K}) \quad (3)$$

Simply, we have

$$\phi(\mathbf{K}) = \sum_{k=1}^n |\sigma_k - s_k| \quad (4)$$

where σ_k stand for prescribed poles, whereas the current dominant ones are expressed as s_k .

Function $\pi(\mathbf{K})$ can be chosen as a penalty function (Fletcher 1987), for instance

$$\pi(\mathbf{K}) = \lambda_1 \alpha_r(\mathbf{K}), \lambda_1 > 0 \quad (5a)$$

$$\pi(\mathbf{K}) = \lambda_2 \max(0, \bar{\alpha}_r - \alpha_r(\mathbf{K}))^l, \quad (5b)$$

$$\bar{\alpha}_r < \min(\sup \text{Re } \sigma_k, \alpha_r(\mathbf{K})), \lambda_2 > 0, l \in \mathbb{N} \geq 1$$

where $\bar{\alpha}_r$ represents the desired eventual abscissa margin for the rest of the spectrum, and λ_2 must be sufficiently high.

Finally, for $\varphi(\mathbf{K})$, a barrier function is better to be taken instead of penalty one, due to the constrain (inequality) in (2), e.g.

$$\varphi(\mathbf{K}) = -\log\left(1 - \varepsilon - \sum_{j=1}^{h_N} |a_{nj}(\mathbf{K})|\right), 0 < \varepsilon \ll 1$$

see (Michiels and Gumussoy 2014), for which the initial setting must be made such that $1 - \varepsilon > \sum_{j=1}^{h_N} |a_{nj}(\mathbf{K}_0)|$.

SIMULATION EXAMPLE

Let us demonstrate the procedure described above on a simulation example in the MATLAB[®]/Simulink[®] environment. Controller structure design is omitted.

Assume a non-minimum phase unstable TDS plant modeled by the transfer function

$$P(s) = \frac{(s-4)e^{-s}}{s+1-2e^{-0.4s}} \quad (6)$$

Although model (6) is of a retarded type, any attempt to design a feasible feedback controller results in the TDS control system of a neutral type. For instance, the use of the stabilizing general finite-dimension controller governed by the transfer function

$$R(s) = \frac{q_2 s^2 + q_1 s + q_0}{s^2 + p_1 s + p_0} \quad (7)$$

within the well-known habitual simple negative feedback loop yields the following characteristic quasipolynomial

$$\begin{aligned} D(s, \mathbf{K}) &= (s+1-2e^{-0.4s}) \left(s^2 + \sum_{i=0}^1 p_i s^i \right) \\ &\quad + ((s-4)e^{-s}) \sum_{i=0}^2 q_i s^i \\ &= (1+q_2 e^{-s}) s^3 \\ &\quad + (1-2e^{-0.4s} + p_1 + (q_1 - 4q_2)e^{-s}) s^2 \\ &\quad + (p_1(1-2e^{-0.4s}) + p_0 + (q_0 - 4q_1)e^{-s}) s \\ &\quad + p_0(1-2e^{-0.4s}) - 4q_0 e^{-s} \end{aligned} \quad (8)$$

the roots of which constitute the system spectrum. Strong stability condition (2) apparently reads

$$|q_2| < 1 \quad (9)$$

which must hold during the evolution of the selectable parameters vector $\mathbf{K} = (p_1, p_0, q_1, q_0, q_2)^T, r = 5$. Following step 2 of Algorithm 1, let the prescribed (desired) poles be represented by a conjugate pair $\sigma_{1,2} = -0.1 \pm 0.2j$, i.e. $n = 2$. Note that it can be calculated that the vertical line introduced in Property 1, item 2, is located in $\gamma = -\log| -1/q_2 | = \log|q_2|$, see e.g. (Bonnet et al. 2011). Hence, in order to get the vertical strip of neutral poles left from the desired pair, the inequality $\log|q_2| < -0.1$ must hold, the solution of which reads

$$|q_2| < 0.9048 \quad (10)$$

By placing roots of (9) directly to the desired loci, the initial feedback spectrum found within the selected range $R = [-6, 15] \times [0, 40]$ and parameters setting read

$$\begin{aligned} \Sigma_0 &= \left\{ \begin{array}{l} 0.1458 \pm 3.94j, -0.1 \pm 0.2j, -0.4731 \pm 22.114j, \\ -0.4737 \pm 16.0155j, -0.5067 \pm 9.709j, \dots \end{array} \right\} \\ \mathbf{K}_0 &= (1.0317, 1.2722, 0.71174, 0.2978, 0.5927)^T \end{aligned}$$

respectively. This setting gives exponentially unstable yet strongly stable feedback system with $\alpha(\mathbf{K}_0) = 0.1458$ and $\gamma = \bar{\gamma} = -0.5231$. It indicates that the spectrum must be enhanced according to step 4 of Algorithm 1.

The evolution of $|\sigma_1 - s_1|$, $\alpha(\mathbf{K})$ and $\alpha_r(\mathbf{K})$ via the QCSA for the number of 16000 iterations can be seen in Figure 1, and that of \mathbf{K} is displayed in Figure 2. As can be seen, the value of q_2 is being improved during the shifting and system has been stabilized. However, the dominant pair is still quite far from the desired one. A slump in the plot of $|\sigma_1 - s_1|$ is caused by closeness of real parts of the dominant pair and a close real pole that appears right from the pair within some iterations.

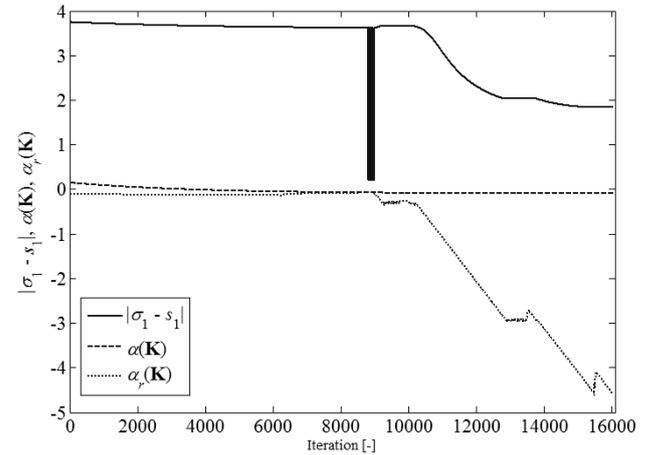


Figure 1: The Evolution of $|\sigma_1 - s_1|$, $\alpha(\mathbf{K})$, $\alpha_r(\mathbf{K})$ by Using the QCSA

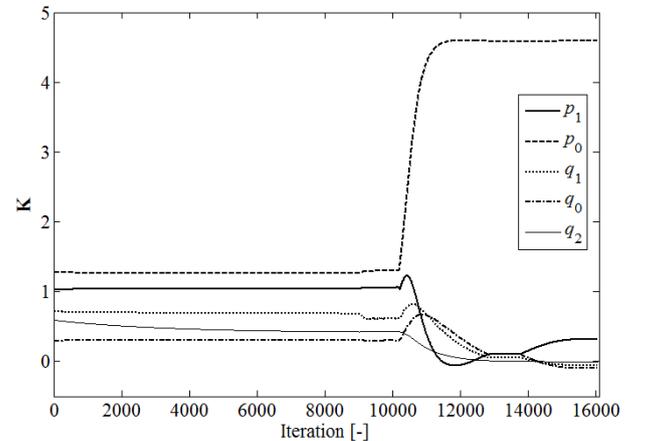


Figure 2: The Evolution of \mathbf{K} by Using the QCSA

The eventual dominant spectrum and the values of \mathbf{K} are the following

$$\Sigma_{16000} = \left\{ \begin{array}{l} -0.0988 \pm 2.0394j, -4.5531, -4.6077, \\ -4.6077, -4.7493, -4.7533 \pm 18.5866j, \dots \end{array} \right\}$$

$$\mathbf{K}_{16000} = (0.316, 4.59, -0.0568, -0.0984, -0.0054)^T$$

These values are taken as initial conditions for the NM optimization procedure (Nelder and Mead 1965), the results of which are provided to the reader in Figures 3 and 4 where Δ stands for the simplex edge length. The objective function has been brought together from (4), (5a) with $\lambda := \lambda_1 = 0.2$ and $\varphi(\mathbf{K}) \equiv 0$.

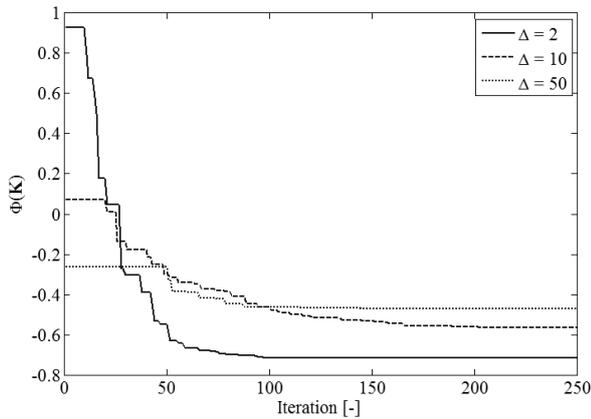


Figure 3: The Evolution of $\Phi(\mathbf{K})$ by Means of the NM Algorithm for $\lambda = 0.2$ and Various Values of Δ

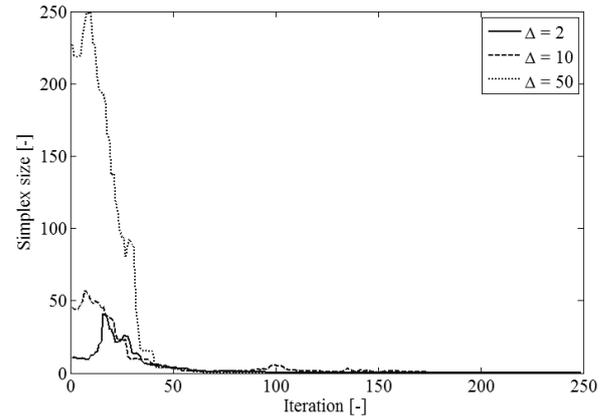


Figure 4: The Evolution of the Simplex Size by Means of the NM Algorithm for $\lambda = 0.2$ and Various Values of Δ

Our simple test has revealed the setting $\lambda = 0.2, \Delta = 2$ as a suitable one for the optimization here since it gives the minimal $\Phi(\mathbf{K})$ from three possibilities. In Figure 5, we further try to compare this setting with the setting pair $\lambda = 0.05, \Delta = 10$. Although results for both pairs in Figure 5 are almost comparable, only the selected setting yields the dominant pair of poles in the close vicinity of desired positions (see the solid thin line approaching the zero).

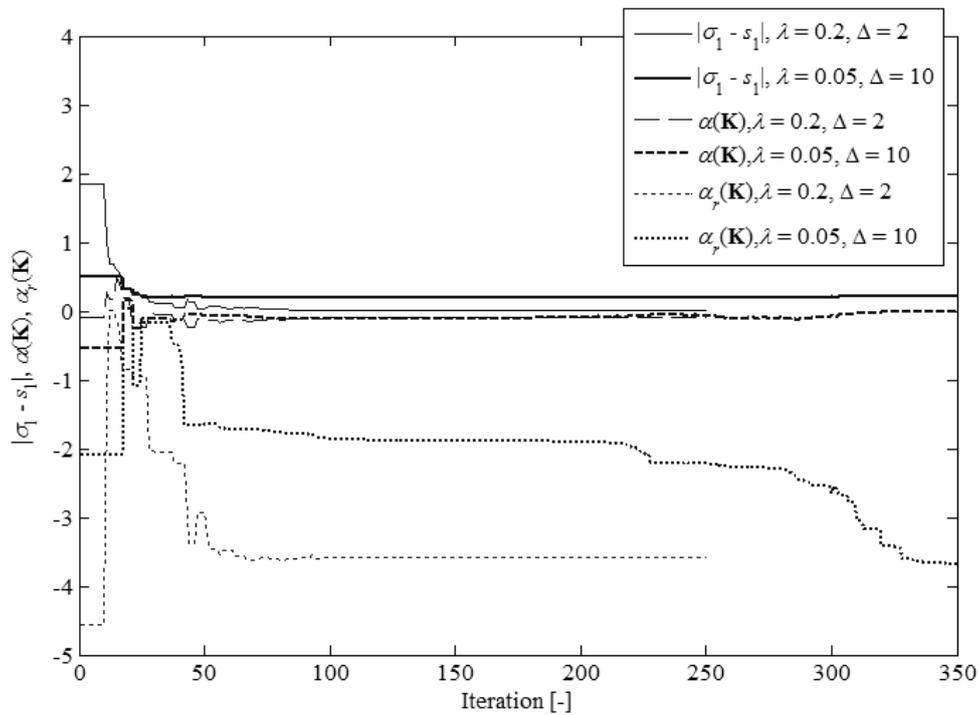


Figure 5: The Evolution of $|\sigma_1 - s_1|$, $\alpha(\mathbf{K})$ and $\alpha_r(\mathbf{K})$ by Means of the NM Algorithm for Setting Pairs $(\lambda, \Delta) = \{(0.05, 10), (0.2, 2)\}$

Then the optimal low-frequency spectrum and the final controller parameters' values read

$$\Sigma_{opt,250} = \left\{ \begin{array}{l} -0.1 \pm 0.2j, -3.59 \pm 4.2255j, \\ -3.5903 \pm 18.9356j, -3.7083 \pm 37.6507j, \\ -3.7579 \pm 31.6205j, -3.9273 \pm 25.0812j, \dots \end{array} \right\}$$

$$\mathbf{K}_{opt,250} = \left(13.4082, -6.3073, -0.3543, -1.7256, \right)^T$$

$$\left(-0.0238 \right)$$

Apparently, the dominant poles agree with desired ones and the feedback system can be judged as strongly exponentially stable (Michiels and Gumussoy 2014) since $\alpha(\mathbf{K}_{opt,250}) = 0.1458$ and condition (9) holds.

The result is also demonstrated by displaying a part of the dominant system spectrum, see Figure 6. Since the rest of the spectrum is quite far from the prescribed pair, it has only a minor effect on the system dynamics.

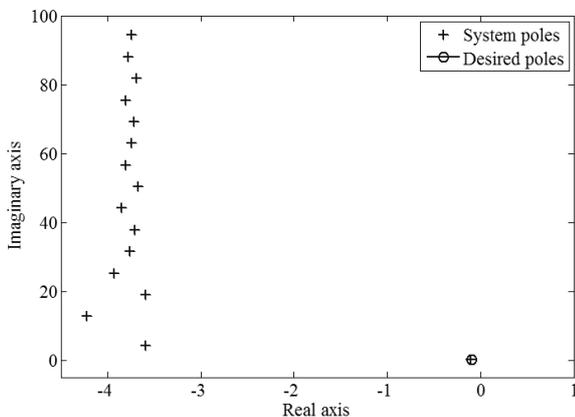


Figure 6: The Eventual Obtained Spectrum

CONCLUSIONS

We have introduced the basic concept of a suboptimal numerical dominant pole assignment procedure for neutral TDS. The goal is to shape the dominant part of the spectrum such that it matches a finite number of prescribed desired poles loci and to push the rest of the spectrum as left as possible. The algorithm consists of three steps: the direct pole placement, successive quasi-continuous shifting and the optimization procedure. The process may stop whenever the desired spectrum is reached. The novelty but also the main drama consists in that neutral TDS are considered. These systems have quite complex spectral and stability issues including the sensitivity to infinitesimally small delays. The method has been verified and demonstrated in the MATLAB[®]/Simulink[®] environment via a simulation example of control of an unstable TDS time delay system (TDS). Some possibilities how to adjust the algorithm are given to the reader as well.

The algorithm can be improved mainly by a more sophisticated optimization; namely, the selection of the

objective function and optimization methods (and its parameters), the use of faster software and hardware tools, or by the development of a better poles loci computation.

ACKNOWLEDGEMENTS

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project no. LO1303 (MSMT-7778/2014).

REFERENCES

- Åström, K.J. and T. Häggglund. 2006. *Advanced in PID Control*. ISA, Research Triangle Park.
- Bonnet, C.; A.R. Fioravanti; and J.R. Partington. 2011. "Stability of Neutral Systems with Commensurate Delays and Poles Asymptotic to the Imaginary Axis." *SIAM Journal on Control and Optimization* 48, No. 2, 498-516.
- Fletcher, R. 1987. *Practical Methods of Optimization*. Wiley, New York.
- Hale, J.K. and S.M. Verduyn Lunel. 1993. *Introduction to Functional Differential Equations*. Springer, New York.
- Chiasson, J. and J.J. Loiseau (Eds.). 2007. *Applications of Time Delay Systems*. Springer, New York.
- Lee, E.B. and S. H. Zak. 1982. "On Spectrum Placement for Linear Time Invariant Delay Systems". *IEEE Transactions on Automation Control* 27, No. 2, 446-449.
- Michiels, W.; K. Engelborghs; P. Vansevevanti; and D. Roose. 2002. "Continuous Pole Placement for Delay Equations". *Automatica* 38, No. 5, 747-761.
- Michiels, W. and T. Vyhlídal. 2005. "An Eigenvalue Based Approach for the Stabilization of Linear Time-Delay Systems of Neutral Type". *Automatica* 41, No. 6, 991-998.
- Michiels, M. and S.-I. Niculescu. 2007. *Stability and Stabilization of Time-Delay Systems. Advances in Design and Control* 12. SIAM, Philadelphia.
- Michiels, W.; T. Vyhlídal; and P. Zitek. 2010. "Control Design for Time-Delay Systems Based on Quasi-Direct Pole Placement". *Journal of Process Control* 20, No.3, 337-343.
- Michiels, W. and S. Gumussoy. 2014. "Eigenvalue-Based Algorithms and Software for the Design of Fixed-Order Stabilizing Controllers for Interconnected Systems with Time-Delays". In *Delay Systems, From Theory to Numerics and Applications, Advances in Delays and Dynamics*. T. Vyhlídal, J-F. Lafay and R. Sipahi (Eds.). Springer, New York, 243-256.
- Nelder, J.A. and R. Mead. 1965. "A Simplex Method for Function Minimization". *The Computer Journal* 7, No. 4, 308-313.
- Pekař, L. and P. Navrátil. 2014. "PPSA: A Tool for Suboptimal Control of Time Delay Systems: Revision and Open Tasks". In *Modern Trends and Techniques in Computer Science, 3rd Computer Science On-line Conference 2014. Advances in Intelligent Systems and Computing* 285. R Šilhavý, R. Šenkeřík, Z. Komínková Oplatková and Z. Prokopová (Eds.). Springer, Heidelberg, 17-28.
- Richard, J.P. 2003. "Time-delay Systems: An Overview of Some Recent Advances and Open Problems". *Automatica* 39, No. 10, 1667-1694.

- Sipahi, R.; T. Vyhlídal; S.-I. Niculescu; and P. Pepe (Eds.). 2012. *Time Delay Systems: Methods, Applications and New Trends*. Springer, New York.
- Vanbiervliet, T.; K. Verheyden; W. Michiels; and S. Vandewalle. 2008. "A Nonsmooth Optimization Approach for the Stabilization of Time-Delay Systems". *ESIAM: Control, Optimisation and Calculus of Variations* 14, No. 3, 478-493.
- Vyhlídal, T. 2003. "Analysis and Synthesis of Time Delay System Spectrum". Doctoral dissertation thesis. Faculty of Mechanical Engineering, Czech Technical University in Prague, Prague.
- Wang, Q.-G.; Z. Zhang; K.J. Åström; and L.S. Chek. 2009. "Guaranteed Dominant Pole Placement with PID Controllers". *Journal of Process Control* 19, No. 2, 349-352.
- Zítek, P. and T. Vyhlídal. 2002. "Dominant Eigenvalue Placement for Time Delay Systems". In *Proceedings of the 5th Portuguese Conference on Automation Control, Control 2002* (Aveiro, Portugal). 605-610.
- Zítek, P.; J. Fišer; and T. Vyhlídal. 2013. "Dominant Three Pole Placement in PID Control Loop with Delay". In *Proceedings of the 9th Asian Control Conference* (Istanbul, Turkey, Jun. 23-26). IEEE, 1-6.

AUTHORS BIOGRAPHY

LIBOR PEKAŘ was born in Zlín, Czech Republic, in 1979. He studied Automation and Control Engineering at Tomas Bata University in Zlín where he obtained his MSc. degree in 2005 and the Ph.D. degree in Technical Cybernetics in 2013.



Nowadays, he has been working as a senior lecturer at the same institute. His research interests are modelling, simulation, autotuning, optimal tuning and algebraic control of time delay systems. His e-mail address is: pekar@fai.utb.cz.

ROMAN PROKOP was born in Hodonín, Czech Republic, in 1952. He obtained his MSc. degree at Czech Technical University in Prague in 1976, then Ph.D. degree at Slovak Technical University in Bratislava in 1983, and he became a professor at Technical University of Brno



in 2004. Now he has been a professor at Tomas Bata University in Zlín. His professional interests cover algebraic control methods, autotuning and robust control. His e-mail address is: prokop@fai.utb.cz.

EVALUATION OF THE PRIMARY METABOLISM OF MONOCULTURES AND YOGHURT STARTERS WITH THE PARTICIPATION OF UREASE-DEFICIENT *STREPTOCOCCUS THERMOPHILUS* STRAINS

Ivan Petelkov*, Rositsa Denkova**, Bogdan Goranov***, Vesela Shopska* Georgi Kostov*, Zapryana Denkova***
*Department "Technology of wine and brewing" ** Department "Biochemistry and molecular biology"; *** Department "Microbiology"
University of Food Technologies, 4002, 26 Maritza boulvd., Plovdiv, Bulgaria

E-mail: george_kostov2@abv.bg; vesi_nevelinova@abv.bg; zdenkova@abv.bg; rositsa_denkova@mail.bg;
i_petelkov92@abv.bg; goranov_chemistry@abv.bg

Nadya Ninova-Nikolova, Zoltan Urshev, Svetlana Minkova
LB Bulgaricum PLC, 12A, Malashevska Str., 1202 Sofia, Bulgaria

E-mail: ninova.n@lbbulgaricum.bg; zurshev@lbbulgaricum.bg; minkova.s@lbbulgaricum.bg

KEYWORDS

Primary metabolism, kinetics, yoghurt starters, urease-deficient, *Streptococcus thermophilus*, modeling

ABSTRACT

The aim of the present work was to evaluate the influence of the lack of urease activity in some *Streptococcus thermophilus* strains on the primary metabolism during lactic acid fermentation. A comparison of the kinetics of the lactic acid process with the participation of urea-utilizing and urease deficient streptococcal strains was performed. It was found that the lack of urease activity in the streptococcal strains had no significant effect on the primary metabolism of the monocultures or the yoghurt starters.

INTRODUCTION

Streptococcus thermophilus strains are important part of the composition of dairy starters. Through their accelerated metabolism they accumulate lactic acid in the media and quickly reduce the pH to ensure the necessary conditions for the growth of the *Lactobacillus delbrueckii* ssp. *bulgaricus* strains. The proven positive symbiotic effect in the growth of *Streptococcus thermophilus* and *Lactobacillus delbrueckii* ssp. *bulgaricus* strains in the composition of yoghurt starters is known as "proto-cooperation" (Angelov et al., 2009; Driesses, 1987).

It is well known that among all the lactic acid bacteria only *Streptococcus thermophilus* strains possess high urease activity and the ability to hydrolyze urea from raw milk. As a result of the hydrolysis urea is decomposed to ammonia and CO₂, which enhances the anaerobic conditions of the medium, thereby stimulating the growth of the *Lactobacillus delbrueckii* ssp. *bulgaricus* strains from the yoghurt starters. At the same time the accumulation of ammonia results in a lower rate of pH reduction, which elongates the fermentation process (Angelov et al., 2009; Arioli et al., 2007; Ninova-Nikolova, 2016).

The main role of *S. thermophilus* in the process of lactic acid fermentation is to accelerate acid formation and accumulation and pH reduction of the medium by the production of lactic acid. Therefore, the rate of acid formation is an important technological parameter as the extension of the acid formation time has a negative impact on product quality and brings negative economic

consequences in the organization of industrial process. The rate of acid formation is a strain-specific metabolic trait that is influenced by different physiological properties such as lactose-galactose metabolism, proteolytic system and urease activity (Ninova-Nikolova, 2016).

Urease activity has a serious negative impact on the quality of dairy products - cheese, yogurt, etc. Therefore, its exclusion from metabolism significantly improves the end-products (Ninova-Nikolova, 2016). Since proto-cooperation between the two types of bacteria is essential in industrial starters, the lack of urease activity must be assessed in relation to possible changes in the primary metabolism and the accumulation of lactic acid. The application of urease-deficient streptococcal strains, result of spontaneous mutation, has practical application due to restrictions on the use of genetically modified organisms (Ninova-Nikolova, 2016).

The purpose of the present work was to study the kinetics of lactic fermentation using urease-deficient streptococcal strains as monocultures and in the composition of yoghurt starters. Thus some of the most common kinetic models to assess the fermentation process were applied.

MATERIALS AND METHODS

Monocultures and yoghurt starters

Symbiotic yoghurt starters and monocultures of *Streptococcus thermophilus* were stored at -196 °C as part of the collection of LB Bulgaricum PLC, Sofia, Bulgaria. The following *Streptococcus thermophilus* monocultures were used: *Streptococcus thermophilus* Ft₃ - isolated from BY LBB 26-12; *Streptococcus thermophilus* Yt₃ - isolated from BY LBB 145-18; *Streptococcus thermophilus* Rzt - isolated from BY LBB Razgrad. After natural selection the following urease-deficient *Streptococcus thermophilus* strains were selected: *Streptococcus thermophilus* Ft₃uD₃ - urease-deficient version of *Streptococcus thermophilus* Ft₃; *Streptococcus thermophilus* Yt₃D₃-1 - urease-deficient version of *Streptococcus thermophilus* Yt₃; *Streptococcus thermophilus* Rzt₄uD₂ - urease-deficient version of *Streptococcus thermophilus* Rzt₄ (Ninova-Nikolova, 2016).

The yoghurt starters used were: Starter LBB BY 26-12; Starter b26 + Ft₃uD₃; Starter LBB BY 145-18; Starter BY 145-Yt₃uD₃-1; Starter LBB BY RAZGRAD; Starter

BY Rzb2 + Rzt4uD2. Each yoghurt starter consisted of a *Lactobacillus delbrueckii* ssp. *bulgaricus* strain and a mutant urease-deficient *Streptococcus thermophilus* strain (Ninova-Nikolova, 2016).

Media

Milk based fermentation media by “LB Bulgaricum” PLC were used in the present work.

Bioreactor, culture conditions and sample analysis

Cultivation was performed in a bioreactor with a working volume of 2 dm³, equipped with a system for monitoring and control of the fermentation process "BIOFLO". The system is equipped with circuits for automatically maintaining the pH. The adjustment of pH was carried out with 2 M NaOH at continuous stirring. The anaerobic conditions of the fermentation were guaranteed by its conduction under inert gas – nitrogen medium. The temperature of the culture medium was maintained automatically through control actions, generated by the control device.

The fermentation process was carried out at temperatures of 39 °C and 43 °C, pH 5.90 and 6.20 at a stirring speed of 150 rpm. After loading the medium and its sterilization at 120 °C for 20 min and subsequent cooling down to the fermentation temperature, it was inoculated with 5% culture of the studied *Streptococcus thermophilus* strain or yoghurt starter. Milk samples were taken by sterile sampling system every 30 minutes since the beginning of the process. They were analyzed to determine the

concentration of viable cells and the titratable acidity. The changes in the acidity of the medium were monitored by the amount of the NaOH used for the neutralization of the lactic acid accumulated in the apparatus.

The total number of viable cells of *Lactobacillus delbrueckii* ssp. *bulgaricus* and *Streptococcus thermophilus* in cfu/cm³ was determined by the pour plate method and the spread plate method in synthetic MRS medium and M₁₇ medium, according to the methodology described in ISO 7889: 2005 (ISO 7889: 2005).

The amount of lactic acid was calculated based of the amount of 2M NaOH used for the neutralization (ISO/TS 11869:2012).

Mathematical models and determination of the kinetic characteristics

The kinetics of the lactic acid fermentation process were examined by the system of differential equations (Birol et al., 1998; Kostov et al., 2012):

$$\begin{aligned} \frac{dX}{dt} &= f(X, S, P) \\ \frac{dP}{dt} &= f(X, S, P) \\ \frac{dS}{dt} &= f(X, S, P) \end{aligned} \quad (1)$$

The used kinetic equations (Birol et al., 1998; Kostov et al., 2012) through which the system (1) acquires a certain type are presented in Table 2.

Table 2

Mathematical models for description of the kinetics of the fermentation process

№	Model	dX/dt	dP/dt	dS/dt
1	Monod	$\mu_{\max} \left(\frac{S}{K_{sx} + S} \right)$	$q_{P\max} \left(\frac{S}{K_{sp} + S} \right) X$	$-\frac{1}{Y_{x/s}} \frac{dX}{dt} - \frac{1}{Y_{p/s}} \frac{dP}{dt}$
2	Tiessier	$\mu_{\max} \left(1 - \exp \left(-\frac{S}{K_{sx}} \right) \right) X$	$q_{P\max} \left(1 - \exp \left(-\frac{S}{K_{sp}} \right) \right) X$	
3	Hinshelwood	$\mu_{\max} \left(\frac{S}{K_{sx} + S} \right) (1 - K_{px} P) X$	$q_{P\max} \left(\frac{S}{K_{sp} + S} \right) (1 - K_{pp} P) X$	
4	Aiba	$\mu_{\max} \left(\frac{S}{K_{sx} + S} \right) \exp(-K_{px} P) X$	$q_{P\max} \left(\frac{S}{K_{sp} + S} \right) \exp(-K_{pp} P) X$	

Parametric identification of the models was carried out in MATLAB environment (Kostov et al., 2012; Mitev and Popova, 1995; Popova 1997). The sum of squared errors of the model output data:

$$F(r) = (X(k_1, \dots, k_n) - X^e)^2 + (S(k_1, \dots, k_n) - S^e)^2 + (P(k_1, \dots, k_n) - P^e)^2 \quad (2)$$

was minimized. For that purpose the function “fmincon” was applied.

The output is vector of “fmincon” are model parameters $k = [k_1, k_2, \dots, k_n]$, where k_1, k_2, \dots, k_n are constants.

The overall differential equations system the function “ode45” was used.

RESULTS AND DISCUSSION

The results of the identification of kinetic models are summarized in Table 3 to Table 6. On Figures 1 and 2 are

presented comparisons of the models for some of monocultures and starter cultures.

Monod-based kinetics

The results of the parameters identification of Monod’s model are presented in Table 3. Data showed that monocultures grew with a good fermentation rate that varies within a relatively narrow range between 0,040 and 0,510 h⁻¹. Only one of the variants was characterized by a higher maximum specific growth rate, but the average rate of the process was comparable to that of the other variants, a fact that will be explained later.

The accumulation of cells of the urease-deficient strains was not affected by the lack of urease in their metabolism. Similar values of the specific growth rate were established in all variants. Consequently, the

absence of the gene responsible for the utilization of urea from the composition of milk did not affect the growth of the cells.

It was typical for streptococcal monocultures that in terms of the accumulation of viable cells they had high affinity to the substrate. The saturation constant K_{SX} for all variants was 0. Only in variant Yt3 no affinity was observed which lowered the maximum specific growth rate and it was within the already commented range between 0,040 and 0,510 h^{-1} .

The accumulation of lactic acid by the streptococcal monocultures proceeded with high specific rate of lactic acid production and with high affinity between the

substrate and the culture. The rate varied between 0,365 and 0,513 $g/(g.h)$, which led to higher economic coefficient that ensured that the substrate was transformed mainly to lactic acid.

All models of the Mono-based kinetics described experimental data with high accuracy, the average error between them and the data was between 4 and 60.

Data showed that the exclusion of the gene responsible for the utilization of urea from milk did not change the primary metabolism of any monoculture. Therefore urease-deficient strains can be successfully included in the composition of yoghurt starters.

Table 3

Kinetic parameters of the growth of monocultures and yoghurt starters in the model of Monod

Monocultures							
	μ_{max}	K_{SX}	q_{pmax}	K_{SP}	$Y_{X/S}$	$Y_{P/S}$	Error*10
Ft3 (pH 5,9/ t 43°C)	0,040	0,000	0,444	0,000	0,376	1,185	5,62
Ft3uD3 (pH 5,9/ t 43°C)	0,048	0,000	0,468	0,000	0,268	1,436	3,00
Yt3 (pH 6,2/ t 43°C)	0,057	0,000	0,513	0,001	0,263	1,542	1,87
Yt3uD3-1 (pH 6,2/ t 43°C)	0,049	0,000	0,500	0,041	1,000	1,022	1,22
Yt3 (pH 6,2/ t 39°C)	0,121	10,000	0,372	0,000	0,713	0,882	3,21
Yt3uD3-1 (pH 6,2/ t 39°C)	0,043	0,000	0,365	0,000	0,237	1,697	4,09
Rzt4 (pH 6,0/ t 43 °C)	0,044	0,000	0,463	0,000	0,224	1,540	4,58
Rzt4uD2 (pH 6,0/ t 43°C)	0,043	0,000	0,474	0,000	0,465	1,140	3,96
Yoghurt starters							
LBB BY 26-12 (medium 1)	0,047	0,000	0,334	0,000	0,300	1,221	2,60
LBB BY 26-12 (medium 2)	0,043	0,000	0,284	0,123	0,350	1,801	1,42
b26+Ft3uD3 (medium 1)	0,077	10,000	0,329	0,000	1,000	0,852	0,51
b26+Ft3uD3 (medium 2)	0,050	0,000	0,233	0,000	0,239	3,636	0,35
LBB BY 145-18 (medium 1)	0,041	3,244	0,256	0,000	0,079	10,000	0,53
LBB BY 145-18 (medium 2)	0,083	1,258	0,244	0,083	0,269	10,000	0,29
BY 145-Yt3uD3-1 (medium 1)	0,035	0,000	0,227	0,000	0,193	1,048	0,68
BY 145-Yt3uD3-1 (medium 2)	0,064	1,496	0,232	0,000	0,214	10,000	0,49
LBB BY RAZGRAD (medium 1)	0,049	0,000	0,349	0,012	0,118	2,853	0,49
LBB BY RAZGRAD (medium 2)	0,083	2,985	0,311	0,564	0,196	10,000	0,4
BY Rzb2+Rzt4uD2 (medium 1)	0,039	0,000	0,289	0,000	0,198	2,255	0,82
BY Rzb2+Rzt4uD2 (medium 2)	0,054	1,095	0,234	0,000	1,000	0,983	0,31

A similar study of the kinetics of the primary metabolism with inclusion of urease-deficient strains in the composition of several yoghurt starters was performed. Due to the specific requirements for the obtaining of the yoghurt starters the cultivation was conducted in two different media.

Data in Table 3 showed that the symbiotic yoghurt starter strains of *Lactobacillus delbrueckii* ssp. *bulgaricus* and *Streptococcus thermophilus* grew with maximum specific growth rate in the range of 0,036-0,083 h^{-1} , but also at higher specific rates of accumulation of lactic acid - 0,284 - 0,349 $g/(g.h)$. The affinity of the strains to the substrate in the medium was high with slight deviations from this summarization concerning only few strains. The inclusion of urease-deficient strains in the composition of yoghurt starters did not alter the primary metabolism; on the contrary, it accelerated the process and milk coagulated 30 min earlier in variants with urease deficient *Streptococcus thermophilus* strain than in the variants in which urea was utilized. This was due to the

previously mentioned fact that the accumulation of ammonia leads to the neutralization of the pH, hence, to slower coagulation.

An interesting fact is that the exclusion of sodium acetate from the composition of the medium caused lack of affinity of the yoghurt starter to the substrate, which was probably due to adversely influencing the utilization of the carbon source and the need for microorganisms to accumulate additional substances which would allow them to utilize lactose from milk.

The main amounts of substrate were utilized and converted to lactic acid, which was confirmed by the higher values of the coefficients $Y_{P/S}$ compared with $Y_{X/S}$ (Table 3).

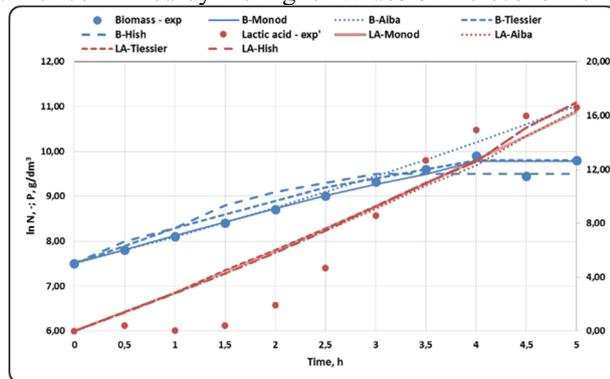
Aiba based kinetics

A common characteristics of many biological processes is that the accumulation of the metabolic products may be associated with the inhibition of growth. The model of Aiba assumes exponential growth inhibition as a result of

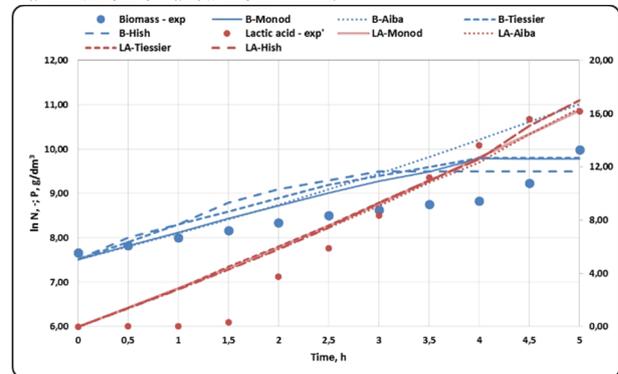
the accumulated lactic acid in the medium. Data from the parameters identification of the model of Aiba are presented in Table 4.

Similarly to Monod's model, the monocultures grew primarily by accumulating lactic acid, which was proven by the high values of the specific rate q_{pmax} and the relatively high values of μ_{max} . This observation was further confirmed by the higher values of the economic

coefficient, which were associated with the accumulation of product obtained from one unit of substrate. The model parameters did not show substrate or product inhibition; the error in describing the test results was minimized. Data from Aiba's model showed that the primary metabolism of streptococcal monocultures was not affected by the exclusion of the gene responsible for the utilization of urea from milk.



a) monoculture of Yt3 (pH 6,2/ t 43°C)



b) monoculture of Yt3uD3-1 (pH 6,2/ t 43°C)

Fig.1. Comparison of kinetic models for cultivation of monoculture of Yt3 and Yt3uD3-1

Table 4

Kinetic parameters of the growth of monocultures and yoghurt starters in the model of Aiba

Monocultures									
	μ_{max}	K_{sx}	q_{pmax}	K_{sp}	$Y_{x/s}$	$Y_{p/s}$	K_{px}	K_{pp}	Error*10
Ft3 (pH 5,9/ t 43°C)	0,066	0,000	0,424	0,000	0,999	1,078	0,000	0,000	14,20
Ft3uD3 (pH 5,9/ t 43°C)	0,073	0,000	0,447	0,000	0,962	1,101	0,000	0,000	11,50
Yt3 (pH 6,2/ t 43°C)	0,108	5,342	0,496	0,001	1,000	1,074	0,000	0,000	5,90
Yt3uD3-1 (pH 6,2/ t 43°C)	0,069	1,468	0,489	0,055	1,000	1,048	0,000	0,000	4,80
Yt3 (pH 6,2/ t 39°C)	0,077	0,000	0,359	0,000	1,000	1,143	0,000	0,000	10,29
Yt3uD3-1 (pH 6,2/ t 39°C)	0,058	0,000	0,355	0,000	0,959	1,084	0,000	0,000	6,20
Rzt4 (pH 6,0/ t 43 °C)	0,083	0,000	0,441	0,000	0,969	1,245	0,000	0,000	17,00
Rzt4uD2 (pH 6,0/ t 43°C)	0,075	0,000	0,449	0,000	0,986	1,209	0,000	0,000	16,40
Yoghurt starters									
LBB BY 26-12 (medium 1)	0,065	0,000	0,323	0,000	0,991	1,033	0,000	0,000	6,90
LBB BY 26-12 (medium 2)	1,425	0,000	0,293	0,428	0,104	0,730	6,231	0,000	2,42
b26+Ft3uD3 (medium 1)	0,061	3,665	0,327	0,000	0,128	10,000	0,000	0,000	3,10
b26+Ft3uD3 (medium 2)	1,500	0,167	0,225	0,000	0,108	0,700	7,051	0,000	11,80
LBB BY 145-18 (medium 1)	0,817	0,000	0,250	0,000	0,257	0,774	4,448	0,000	2,87
LBB BY 145-18 (medium 2)	0,085	0,789	0,243	0,010	1,000	1,120	0,000	0,000	0,59
BY 145-Yt3uD3-1 (medium 1)	0,051	0,000	0,224	0,000	0,107	10,000	0,071	0,000	4,41
BY 145-Yt3uD3-1 (medium 2)	0,056	0,210	0,236	0,001	0,702	1,173	0,000	0,000	1,21
LBB BY RAZGRAD (medium 1)	0,871	0,000	0,336	0,000	0,470	0,767	2,470	0,000	3,50
LBB BY RAZGRAD (medium 2)	1,354	0,483	0,281	0,002	1,000	0,953	5,863	0,000	1,71
BY Rzb2+Rzt4uD2 (medium 1)	0,052	0,000	0,283	0,000	0,990	1,058	0,000	0,000	3,38
BY Rzb2+Rzt4uD2 (medium 2)	0,049	0,000	0,234	0,000	0,994	1,031	0,000	0,000	0,88

Using the model of Aiba to describe the kinetics of the growth of symbiotic yoghurt starters showed other trends in comparison to those obtained by Monod's model. The maximum specific growth rate for some variants reached values between 1,35 and 1,50 h^{-1} , but at the same time substrate inhibition in these variants was observed. Thus, the overall growth rate would not be different from the other variants. Substrate inhibition was observed with both types of media, which was most likely due to the

excess of some components of the medium. There was no product inhibition of the process of accumulation of lactic acid - the inhibition constant K_{pp} equaled 0 for all variants.

The inclusion of urease-deficient *Streptococcus thermophilus* strain in the composition of the yoghurt starters did not alter the primary metabolism of the yoghurt starter. Data from Aiba's model showed slight

increase in the specific growth rate of the symbiotic culture.

Hinshelwood based kinetics

Hinshelwood's model assumes a non-linear growth inhibition and the accumulation of a product which unlike the exponential one, is not as strong. Data from the identification of the parameters are presented in Table 5. Data from Hinshelwood's model, in accordance with the other two tested models, showed that monocultures were growing with maximum specific growth rate at maximum affinity to the substrate. The main amount of the substrate was utilized for the accumulation of lactic acid, the rate of this process being 8-10 times higher than that for the accumulation of biomass. This was confirmed

by the values of the economic coefficient $Y_{P/S}$. The metabolism of the cells was not affected by the lack of urease in the mutants.

There were significant differences in studying the yoghurt starters in comparison with the conclusions drawn from the previous models. Minor to moderate lack of affinity to the substrate was typical for some of them. It might be provoked by the observed substrate inhibition of the growth of the symbiotic culture. Again, there was no product inhibition.

The error of the model as compared to the experimental data was minimal. The addition of the urease-deficient strain in the composition of the yoghurt starter did not change substantially the primary metabolism.

Table 5

Kinetic parameters of the growth of monocultures and yoghurt starters according to the model of Hinshelwood

	μ_{max}	K_{SX}	q_{pmax}	K_{SP}	$Y_{X/S}$	$Y_{P/S}$	K_{PX}	K_{PP}	Error*10
Monocultures									
Ft3 (pH 5,9/ t 43°C)	0,065	0,000	0,423	0,000	0,153	13,977	0,000	0,000	11,400
Ft3uD3 (pH 5,9/ t 43°C)	0,073	0,000	0,447	0,000	0,161	16,188	0,000	0,000	11,500
Yt3 (pH 6,2/ t 43°C)	0,075	0,158	0,495	0,000	1,144	1,046	0,000	0,000	5,850
Yt3uD3-1 (pH 6,2/ t 43°C)	0,063	0,388	0,487	0,003	0,189	2,344	0,000	0,000	4,810
Yt3 (pH 6,2/ t 39°C)	0,178	16,197	0,352	0,000	100,000	0,919	0,000	0,000	10,150
Yt3uD3-1 (pH 6,2/ t 39°C)	0,058	0,000	0,355	0,000	7,525	0,934	0,000	0,000	6,180
Rzt4 (pH 6,0/ t 43 °C)	0,076	0,000	0,438	0,000	0,214	3,759	0,000	0,000	15,210
Rzt4uD2 (pH 6,0/ t 43°C)	0,073	0,000	0,450	0,000	0,469	1,380	0,000	0,000	14,520
Yoghurt starters									
LBB BY 26-12 (medium 1)	0,064	0,000	0,324	0,000	0,205	3,154	0,000	0,000	7,310
LBB BY 26-12 (medium 2)	0,096	6,145	0,279	0,101	100,000	0,999	0,000	0,000	2,860
b26+Ft3uD3 (medium 1)	0,091	0,000	0,321	0,000	1,763	0,873	0,100	0,000	3,150
b26+Ft3uD3 (medium 2)	0,059	0,300	0,231	0,000	27,928	0,853	0,000	0,000	1,650
LBB BY 145-18 (medium 1)	0,192	50,000	0,253	0,000	0,089	5,511	0,024	0,000	2,800
LBB BY 145-18 (medium 2)	0,499	37,470	0,239	0,016	4,216	0,894	0,000	0,000	0,632
BY 145-Yt3uD3-1 (medium 1)	0,049	0,000	0,224	0,000	0,101	100,000	0,051	0,000	4,450
BY 145-Yt3uD3-1 (medium 2)	0,470	47,102	0,231	0,001	2,025	0,969	0,000	0,001	1,280
LBB BY RAZGRAD (medium 1)	0,070	0,948	0,343	0,000	0,103	100,000	0,028	0,000	2,280
LBB BY RAZGRAD (medium 2)	0,060	0,006	0,290	0,035	0,324	1,707	0,029	0,000	1,260
BY Rzb2+Rzt4uD2 (medium 1)	0,052	0,000	0,283	0,000	1,515	0,991	0,000	0,000	3,810
BY Rzb2+Rzt4uD2 (medium 2)	0,053	0,419	0,233	0,000	2,809	0,905	0,000	0,000	0,880

Tiessier based kinetics

Tiessier's models are modified Monod's models, but they suggest exponential reduction of the specific rates due to substrate inhibition. The results of the parameters identification are presented in Table 6.

The results of this model were highly contradictory. On one hand, the model describes satisfactorily part of the experimental data. With the other part of the data an error was observed, especially when describing the amount of the utilized substrate. On the other hand, the model showed that part of the monocultures had no affinity to the substrate, but, quite unexpectedly, a trend associated with urease deficit was not observed. On the contrary, some of the strains having urease activity, showed no affinity to the substrate. The explanation might be sought in the cultivation conditions. All monocultures grew at

low specific growth rate, while those which had high values of the maximum specific growth rate showed no affinity to the substrate.

All monocultures showed high specific rate of accumulation of lactic acid. Data indicated that the substrate was utilized mainly for the production and accumulation of lactic acid.

As with the other models there was no effect on the primary metabolism in urease-deficient monocultures and yoghurt starters.

The description of the lactic acid process in the symbiotic yoghurt starter also showed interesting tendencies. For example, for almost all yoghurt starters medium 2 proved to be unsuitable for the conduction of the fermentation process. In this medium there was a significant lack of affinity to the substrate, wherein the overall growth rate decreased. The model of Tiessier gave conflicting

information about the metabolism in the yoghurt starters but it could be concluded that the primary metabolism was not affected.

Tiessier's model showed that changes related to the inclusion of urease-deficient streptococcal strains in the composition of the yoghurt starters can be observed, but this would be a subject of further investigation.

CONCLUSION

It was found that the primary metabolism of monocultures of urease-deficient streptococci and yoghurt starters with the inclusion of urease-deficient

strains of *Streptococcus thermophilus* was not affected by the lack of urease activity. All tested models showed that the lactic acid fermentation process was characterized by good growth rates and very high values of the specific rates of accumulation of lactic acid. The process of lactic acid fermentation was not associated with substrate or product inhibition, except for few variants, but the inhibition in these variants was weak.

The results obtained with Tiessier's model, which showed lack of affinity of some monocultures and yoghurt starters to the culture medium, are interesting and give the possible ground for future research.

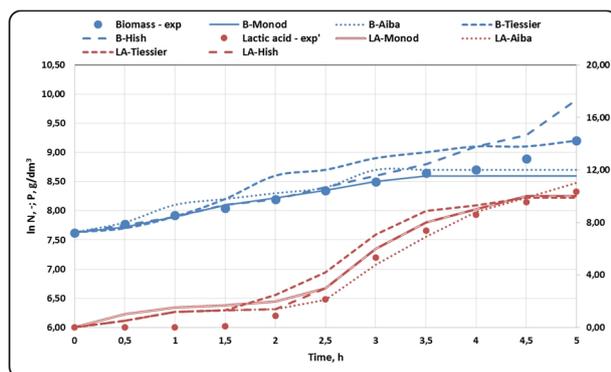
Table 6

Kinetic parameters of the growth of monocultures and yoghurt starters according to the model of Tiessier

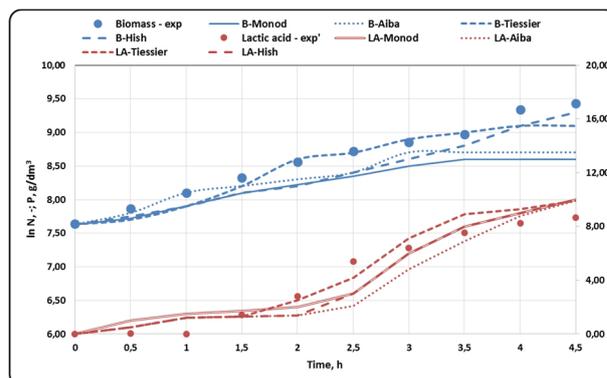
	μ_{max}	K _{sx}	q _{pmax}	K _{sp}	Y _{x/s}	Y _{p/s}	Error*10
Monocultures							
6Ft3 (pH 5,9/ t 43°C)	0,055	1,608	0,432	0,094	6,000	0,940	11,600
Ft3uD3 (pH 5,9/ t 43°C)	0,063	1,945	0,455	0,076	6,000	0,946	11,700
Yt3 (pH 6,2/ t 43°C)	0,061	2,266	0,500	0,269	100,000	0,933	6,110
Yt3uD3-1 (pH 6,2/ t 43°C)	0,688	156,419	0,489	0,952	100,000	0,930	4,810
Yt3 (pH 6,2/ t 39°C)	0,107	11,585	0,358	0,014	95,693	0,919	10,300
Yt3uD3-1 (pH 6,2/ t 39°C)	0,051	0,000	0,360	0,007	0,955	1,060	6,310
Rzt4 (pH 6,0/ t 43 °C)	0,714	147,222	0,442	0,137	98,697	1,006	11,800
Rzt4uD2 (pH 6,0/ t 43°C)	0,201	202,500	0,455	0,269	98,697	1,000	12,100
Yoghurt starters							
LBB BY 26-12 (medium 1)	0,056	0,335	0,329	0,092	0,649	0,979	7,410
LBB BY 26-12 (medium 2)	1,813	250,000	0,278	0,884	66,830	0,998	2,800
b26+Ft3uD3 (medium 1)	0,270	75,560	0,326	0,019	0,685	0,974	3,180
b26+Ft3uD3 (medium 2)	1,867	226,463	0,229	0,001	217,022	0,850	1,720
LBB BY 145-18 (medium 1)	0,692	250,000	0,254	0,805	0,083	7,090	2,380
LBB BY 145-18 (medium 2)	1,494	131,465	0,242	0,657	29,876	0,865	0,630
BY 145-Yt3uD3-1 (medium 1)	0,373	122,830	0,224	0,027	0,185	1,163	4,480
BY 145-Yt3uD3-1 (medium 2)	0,058	1,934	0,237	0,363	39,202	0,862	1,200
LBB BY PA3ΓPAΔ (medium 1)	0,475	100,872	0,342	0,001	0,229	1,110	3,510
LBB BY PA3ΓPAΔ (medium 2)	2,377	250,000	0,290	0,661	69,911	0,867	1,220
BY Rzb2+Rzt4uD2 (medium 1)	0,039	1,035	0,000	249,643	13,802	5,802	1,170
BY Rzb2+Rzt4uD2 (medium 2)	1,209	227,701	3,995	0,086	16,415	37,806	1,120

REFERENCES

- Angelov M., Kostov G., Simova E., Beshkova D., Koprinkova-Hristova P., 2009. Proto-cooperation factors in yogurt starter culture." <http://www.revue-genie-industriel.info/document.php?id=755>, 3 (2009) ISSN 1313-8871
- Arioli S., Monnet C., Guglielmetti S., Parini C., De Noni I., Hogenboom J., Halami P.M., Mora D., 2007. "Aspartat Biosynthesis Is Essential for the Growth of *Streptococcus thermophilus* in Milk, and Aspartat Availability Modulates the Level of Urease activity." *Applied and Environmental microbiology* 73(18), 5789-5796.
- Birol G., Doruker P., Kardar B., Onsan Z., Ulgen K. 1998. "Mathematical description of ethanol fermentation by immobilised *Saccharomyces cerevisiae*", *Process Biochemistry*, 33, 763-771.
- Driessen F.M. 1987. "Protocooperation of yogurt bacteria in continuous culture", In: Mixed Cultures Fermentation, ed by Buchell ME, Slater J.H. Academic Press, New York, pp. 99-120.
- 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
- 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.
- Ninova-Nikolova, N., 2016. "Selection of urease deficit strains *Streptococcus thermophilus* and their application for fermented milk food." Thesis, UFT, Plovdiv, p.176.
- 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
- ISO 7889:2003 (IDF 117:2003) - Yogurt -- Enumeration of characteristic microorganisms -Colony-count technique at 37 degrees C
- ISO/TS 11869:2012 (IDF 150) - Fermented milks -- Determination of titratable acidity -- Potentiometric method



a) medium A



b) medium B

Fig.2. Comparison of kinetic models for cultivation of starter cultures with contribution of Yt3uD3-1

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, PhD on “Mechanical engineering in food and flavor industry (Technological equipment in biotechnology industry)” in 2007 from University of Food Technologies, Plovdiv and DSc on “Intensification of fermentation processes with immobilized biocatalysts”. His research interests are in the area of bioreactors construction, biotechnology, microbial population’s investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, beer production.

VESELA SHOPSKA is assistant professor 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 2006 at University of Food Technologies, Plovdiv. She received her PhD in “Technology of alcoholic and non-alcoholic beverages (Brewing technology)” in 2014. Her research interests are in the area of beer fermentation with free and immobilized cells; yeast and bacteria metabolism and fermentation activity.

ZAPRYANA DENKOVA is professor at the department “Microbiology” at University of Food Technologies, Plovdiv. She received her MSc in “Technology of microbial products” in 1982, PhD in „Technology of biologically active substances“ in 1994 and DSc on “Production and application of probiotics” in 2006. Her research interests are in the area of selection of probiotic strains and development of starters for food production, genetics of microorganisms, and development of functional foods.

ROSITSA DENKOVA is assistant professor at the department “Biochemistry and molecular biology” at University of Food Technologies, Plovdiv. She received her MSc in “Industrial biotechnologies” in 2011 and PhD in “Biotechnology (Technology of biologically active substances)” in 2014. Her research interests are in the area of isolation, identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is researcher at the department “Microbiology” at University of Food Technologies, Plovdiv. He received his PhD in 2015 from University of Food Technologies, Plovdiv. The theme of his thesis was “Production of lactic acid with free and immobilized lactic acid bacteria and its application in food industry”. His research interests are in the area of bioreactors construction, biotechnology, microbial population’s investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, beer production.

NADYA NINOVA-NIKOLOVA is researcher at LB Bulgaricum PLC, “Starter cultures” laboratory. She received her PhD in 2016 from University of Food Technologies, Plovdiv. The theme of her thesis was “Selection of urease-deficient strains *Streptococcus thermophilus* and their application for fermented milk products”. Her research interests are in the area of fermentation of lactic acid bacteria and probiotic strains in bioreactors, development of starter cultures for yoghurt, cheese, butter, bread.

ZOLTAN URSHEV is researcher at LB Bulgaricum PLC, “DNA-analysis” laboratory, PhD in Microbiology since 2009.

SVETLANA MINKOVA is associated professor at LB Bulgaricum PLC. She received her PhD degree in 2001 at University of Food technologies – Plovdiv in the field of dairy sciences. Now she is the Department Director R&D and Licensing at LB Bulgaricum Plc.

MODELING AND ANALYSIS OF SPIN SPLITTING IN STRAINED GRAPHENE NANORIBBONS

Sanjay Prabhakar,¹ Roderick Melnik,^{1,2} and Luis Bonilla³

¹The MS2Discovery Interdisciplinary Research Institute, M2NET Laboratory,
Wilfrid Laurier University, Waterloo, ON, N2L 3C5 Canada

²BCAM-Basque Center for Applied Mathematics, E48009 Bilbao, Spain

³Gregorio Millan Institute, Fluid Dynamics, Nanoscience and Industrial Mathematics,
Universidad Carlos III de Madrid, 28911, Leganes, Spain

KEYWORDS

Complex systems, graphene, coupled theory, Dirac energy bands, multiscale effects.

ABSTRACT

We study the influence of ripple waves, originating from the electromechanical effects, on band structures of graphene nanoribbons (GNRs). GNRs are complex systems that require novel approaches for their analysis, due to multiscale and multiphysics effects involved. Here, we develop a mathematical model and we show that the externally applied magnetic fields along z-direction in combination with pseudo-fields enhance the spin splitting of GNRs bands. In particular, we show that the strain tensor induce quantum confinement effect that turn to lead the opening of the bandgaps at Dirac point. Such finite band gaps are highly sensitive to the control parameters (period length, applied stress) of the ripple waves that help to design the optoelectronic devices for straintronic and spintronic applications.

INTRODUCTION

Graphene has a potential interest for future optoelectronic devices due to its unique electronic and physical properties (Sarma et al. 2011, Castro et al. 2009, Novoselov et al. 2005, Abanin et al. 2006, Barbier et al. 2006). Several observed quantum properties such as the half integer quantum Hall effect, non-zero Berry phase, as well as the measurement of conductivity of electrons in the electronic devices lead to novel applications in carbon based nanoelectronic devices (Sarma et al. 2011, Castro et al. 2009, Novoselov et al. 2005, Novoselov and Jiang et al. 2005, Novoselov et al. 2004, Zhang et al. 2005). One atom thick graphene sheet has the same properties as a two dimensional system that does not contain any band gap at two Dirac points (Novoselov et al. 2005). Further strain engineering of graphene by controlling the electromechanical properties via the pseudomorphic gauge fields are considered as a next generation optoelectronic devices (Shenoy et al. 2008, Choi et al. 2010, Bao et al. 2012, Cadelano et al. 2009, Bao et al.

2009). Small band gap opening is also expected due to implementation of strain tensor in the band structures of graphene through pseudomorphic vector potential.

Due to a range of multiscale effects associated with these properties, the development of new mathematical models and their efficient computational implementations are essential. In this paper we present a model that couples the Navier equations, accounting for electromechanical effects, to the electronic properties of zigzag graphene nanoribbons. We show that the ripple waves originating from the electromechanical effects strongly influence the band structures of GNRs. This response mechanism might be used for tuning the band gaps at the Dirac point in strained GNRs that can be utilized to design the optoelectronic devices for the application in straintronic (Levy et al. 2010).

Experimental studies on two dimensional images of graphene sheet taken from high resolution transmission electron microscope or scanning tunneling microscope show that in-plane and out-of-plane ripples waves varies by several degrees and reach to the nanometer scale (Sarma et al. 2011). These ripples in graphene are induced by several different mechanisms that have been widely investigated (Shenoy et al. 2008, Cadelano et al. 2009, Choi et al. 2010, Kitt et al. 2012, Carpio and Bonilla 2008, Cadelano and Colombo 2012). Such ripples are part of the intrinsic properties of graphene that are expected to strongly affect the band structures due to their coupling through pseudomorphic vector potential (Bao et al. 2009, Bao et al. 2012, Cerda and Mahadevan 2003). Recently in Ref. (Prabhakar et al. 2014), in-plane oscillations and thermomechanics of relaxed-shape graphene due to externally applied tensile edge stress along both the armchair and zigzag directions in graphene quantum dots have also been explored. Here authors have shown that the level crossing between the ground and first-excited states in the localized edge states can be achieved with accessible values of temperature. The level crossing is absent in the states formed at the center of the graphene sheet due to the presence of threefold symmetry. More recently, in Ref. (Prabhakar and Melnik 2015), the

authors of this work have investigated the influence of in-plane ripple waves in graphene nanoribbons. In this paper, we investigate the influence of both in-plane and out-of-plane ripple waves in zigzag graphene nanoribbons in presence of external magnetic fields along z-direction. In particular we present a model that couples the Navier equations, accounting for electromechanical effects, to the electronic properties of zigzag graphene nanoribbons. We show that the ripple waves originating from the electromechanical effects strongly influence the band structures of GNRs. This response mechanism might be used for tuning the band gaps at the Dirac point in strained GNRs that can be utilized to design the optoelectronic devices for the application in straintronics. Other novel applications of complex systems, such as GNRs, based on coupled physical effects are expected.

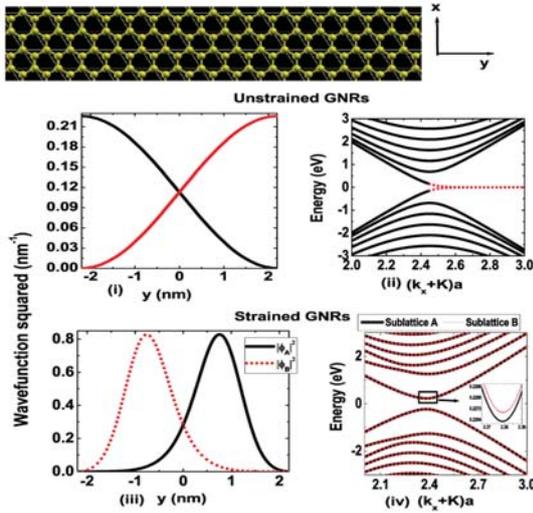


FIGURE 1: Schematics of zigzag graphene nanoribbons is shown in upper panel. (a) Ground state wavefunctions squared vs distance (y) in zigzag GNRs at $k_x=0$. These plots are obtained from Eqs. (23) and (24). (b) The band structures of GNR with the zigzag edge are obtained from Eqs.(19) (dotted lines) and (22) (solid lines). Here we choose the width of the GNR as $L=3\sqrt{3}$ aN ($N=6$) to reproduce the results of Ref. (Brey and Fertig 2006). Band diagram of strained zigzag GNRs are shown in Fig. 1 (iii) and (iv). Here we clearly see the band gap openings due to strain tensor and highly asymmetric wavefunctions of graphene electrons of sublattices A and B are observed. The parameters are chosen as: $\tau_e = -100\text{eV/nm}$, $h_0=1\text{nm}$ and $B=1\text{T}$.

MATHEMATICAL MODEL

The total elastic energy density associated with the strain for the two dimensional graphene sheet can be written as (Landau and Lifshitz 1970, Carpio and Bonilla 2008, Cadelano and Colombo 2012) $2U_s = C_{iklm}\epsilon_{ik}\epsilon_{lm}$. Here C_{iklm} is a tensor of rank four

(the elastic modulus tensor) and ϵ_{ik} or ϵ_{lm} is the strain tensor. In the above, the strain tensor components can be written as

$$\epsilon_{ik} = \frac{1}{2}(\partial_{x_k} u_i + \partial_{x_i} u_k + \partial_{x_k} h \partial_{x_i} h), \quad (1)$$

where u_i and h are in-plane and out-of-plane displacements, respectively (Juan et. al. 2013, Carpio and Bonilla 2008, Bao et al. 2009, Shenoy et al. 2008, Cerda and Mahadevan 2003). Hence, the strain tensor components for graphene in the 2D displacement vector $u(x, y) = (u_x, u_y)$ can be written as

$$\epsilon_{xx} = \partial_x u_x + \frac{1}{2}(\partial_x h)^2, \quad (2)$$

$$\epsilon_{yy} = \partial_y u_y + \frac{1}{2}(\partial_y h)^2, \quad (3)$$

$$\epsilon_{xy} = \frac{1}{2}(\partial_y u_x + \partial_x u_y) + \frac{1}{2}(\partial_x h)(\partial_x h). \quad (4)$$

The stress tensor components $\sigma_{ik} = \partial U_s / \partial \epsilon_{ik}$ for graphene can be written as

$$\sigma_{xx} = C_{11}\epsilon_{xx} + C_{12}\epsilon_{yy}, \quad (5)$$

$$\sigma_{yy} = C_{12}\epsilon_{xx} + C_{22}\epsilon_{yy}, \quad (6)$$

$$\sigma_{xy} = 2C_{66}\epsilon_{xy}. \quad (7)$$

In the continuum limit, elastic deformations of graphene sheets under applied tensions are described by the Navier equations $\partial_j \sigma_{ik} + F_i / t = 0$, where F_i are applied tensions. Hence, the coupled Navier equations of electroelasticity can be written as: (Landau and Lifshitz 1970)

$$\begin{aligned} & (C_{11}\partial_x^2 + C_{66}\partial_y^2)u_x + (C_{12} + C_{66})\partial_x\partial_y u_y \\ & + \frac{1}{2}\partial_x [C_{11}(\partial_x h)^2 + C_{12}(\partial_y h)^2] \\ & + C_{66}\partial_y(\partial_x h)(\partial_y h) + \frac{F_x}{t}, \end{aligned} \quad (8)$$

$$\begin{aligned} & (C_{66}\partial_x^2 + C_{11}\partial_y^2)u_y + (C_{12} + C_{66})\partial_x\partial_y u_x \\ & + \frac{1}{2}\partial_y [C_{12}(\partial_x h)^2 + C_{22}(\partial_y h)^2] \\ & + C_{66}\partial_x(\partial_x h)(\partial_y h) + \frac{F_y}{t}, \end{aligned} \quad (9)$$

where t is the thickness of the single layer graphene, $F_x = \tau_e q \sin(qx)$ and $F_y = \tau_e q \sin(qy)$. Here $q = 2\pi / \iota$ with ι being the period length of the in-plane ripple waves. We assume symmetric out-of-plane ripple waves, $\partial_x h = kh_0 \cos(kx)$,

$\partial_y h = kh_0 \cos(ky)$, where $k = 2\pi/l$, l is the period and h_0 is the height of out-of-plane ripple waves) travel along x and y direction in the plane of two dimensional graphene sheet (Meng et al. 2013, Guinea et al. 2008, Bao et al. 2009). Thus, we write Eqs.(8) and (9) as:

$$\begin{aligned} & (C_{11}\partial_x^2 + C_{66}\partial_y^2)u_x + (C_{12} + C_{66})\partial_x\partial_y u_y \\ &= \frac{1}{2}C_{11}k^3h_0^2 \sin(2kx) + C_{66}k^3h_0^2 \cos(kx)\sin(ky) \\ & - \frac{\tau_e q}{t} \sin(qx), \end{aligned} \quad (10)$$

$$\begin{aligned} & (C_{66}\partial_x^2 + C_{11}\partial_y^2)u_y + (C_{12} + C_{66})\partial_x\partial_y u_x \\ &= \frac{1}{2}C_{22}k^3h_0^2 \sin(2ky) + C_{66}k^3h_0^2 \sin(kx)\cos(ky) \\ & - \frac{\tau_e q}{t} \sin(qy), \end{aligned} \quad (11)$$

For zigzag GNRs elongated along x-direction and applying tensile edge stress along y-direction, we assume ε_{yy} is non-vanishing strain tensor component. Therefore, from Eq (9), we write the strain tensor as (Meng 2013):

$$\varepsilon_{yy} = \frac{\tau_e}{C_{22}t} \cos(qy) + \frac{1}{4}k^2h_0^2 + \frac{kh_0^2}{4L} \sin(kL) - \frac{2\tau_e}{qC_{22}tL} \sin\left(\frac{qL}{2}\right). \quad (12)$$

Now we turn to the influence of strain tensor on the electronic properties of zigzag graphene quantum dots.

In the continuum limit, by expanding the momentum close to the K point in the Brillouin zone, the Hamiltonian for π electrons at the K point reads as (Maksym and Aoki 2013, Krueckl and Richter 2012, Neto et al. 2009):

$$H = v_F(\sigma_x P_x + \sigma_y P_y) + \frac{1}{2}g_0\mu_B B\sigma_z, \quad (13)$$

In (13), $P = p - \eta A_s - eA$ with $p = -i\hbar\partial_x$ being the canonical momentum operator, $A_s = (-\varepsilon_{yy}, 0)\beta/a$ is the vector potential induced by pseudomorphic strain tensor and $A = B(-y, 0)$ is the vector potential due to applied magnetic field, B along z-direction (Kitt et al. 2012, Guinea et al. 2008, Guinea and Horovitz et al. 2008, Juan et al. 2013). The last term is the Zeeman energy.

For strained graphene nanoribbons with zigzag edge (Sevincli et al. 2008, Zheng et al. 2007), we assume $H\psi = \varepsilon\psi$, where $\psi(r) = \exp(ik_x x)(\phi_A(y), \phi_B(y))^T$ (Neto et al. 2009). Thus the two coupled equations can be written as

$$\left(k_x - \partial_y + \frac{\beta}{a}\varepsilon_{xx} + \frac{eB}{\eta}y\right)\phi_B = \left(\frac{\varepsilon - g_0\mu_B B/2}{\eta v_F}\right)\phi_A, \quad (14)$$

$$\left(k_x + \partial_y + \frac{\beta}{a}\varepsilon_{xx} + \frac{eB}{\eta}y\right)\phi_B = \left(\frac{\varepsilon + g_0\mu_B B/2}{\eta v_F}\right)\phi_A. \quad (15)$$

In the general case, exact solutions of (14) and (15) are not feasible to find. However for some special cases, e.g.

for unstrained Hamiltonians without any source of external magnetic field, we can write two coupled equations as:

$$(k_x - \partial_y)\phi_B = \left(\frac{\varepsilon}{\eta v_F}\right)\phi_A, \quad (16)$$

$$(k_x + \partial_y)\phi_A = \left(\frac{\varepsilon}{\eta v_F}\right)\phi_B, \quad (17)$$

Now, we can apply operator $(k_x + \partial_y)$ to Eq.(16) and by using Eq.(17), we can arrive at the second order partial differential equation:

$$(\eta v_F)^2(k_x^2 - \partial_y^2)\phi_B = \varepsilon^2\phi_B. \quad (18)$$

The unstrained GNRs with zigzag edges support two different states such as surface waves (edge states) which exist at or near the edge and confined modes. Thus, we write the energy spectrum of the nanoribbons near the edge as:

$$\varepsilon_{n\pm}^z = \pm\sqrt{(\eta v_F)^2(k_x^2 - z^2)}, \quad (19)$$

where z is a real number which follows the solution of:

$$\exp(-2zL) = (k_x - z)/(k_x + z). \quad (20)$$

For $z = ik_n$, the transcendental Eq.(20) becomes

$$k_x = k_n \cot(k_n L), \quad (21)$$

and the energy spectrum of GNR for confined modes is given by

$$\varepsilon_{n\pm}^{zc} = \pm\sqrt{(\eta v_F)^2(k_x^2 + k_n^2)}, \quad (22)$$

Also the wavefunctions $\phi_A(y)$ and $\phi_B(y)$ for GNR with zigzag edge is given by

$$\phi_A = \frac{2iN}{\varepsilon_{n\pm}^{zc}} \{k_x \sin(k_n y) - k_n \cos(k_n y)\}, \quad (23)$$

$$\phi_B = 2iN \sin(k_n y). \quad (24)$$

Since the wavefunctions do not admit the valleys for zigzag GNR, we assume $\langle \phi'_A(y) | \phi'_A(y) \rangle = \langle \phi'_B(y) | \phi'_B(y) \rangle = 0$ and apply the normalization condition $\langle \phi_A(y) | \phi_A(y) \rangle + \langle \phi_B(y) | \phi_B(y) \rangle = 1$ to find the constant N as:

$$|N|^2 = \frac{k_n (\varepsilon_{n\pm}^{zc})^2}{\tilde{\kappa} + 4k_n (k_n^2 L - k_x \sin^2(k_n L))}, \quad (25)$$

where

$$\tilde{\kappa} = (k_x^2 - k_n^2 + (\varepsilon_{n\pm}^{zc})^2) \{2k_n L - \sin(2k_n L)\} \quad (26)$$

At $k_x = 0$, Eqs.(23) and (24) can be written as

$$\phi_A = \mu \frac{i}{\sqrt{L}} \cos\left[\frac{(2n+1)\pi y}{2L}\right], \quad (27)$$

$$\phi_B = \mu \frac{i}{\sqrt{L}} \sin\left[\frac{(2n+1)\pi y}{2L}\right], \quad (28)$$

Where $n = 0, 1, 2, \dots$. The wavefunctions and eigenvalues of the unstrained zigzag GNRs are shown in Figs. 1(i) and (ii).

For strained GNRs, we can apply the operator

$$\left(k_x + \partial_y + \frac{\beta}{a}\varepsilon_{xx} + \frac{eB}{\eta}y\right)$$

from left on (14) and the operator

$$\left(k_x - \partial_y + \frac{\beta}{a}\varepsilon_{xx} + \frac{eB}{\eta}y\right)$$

from left on (15) and cast these two coupled Eqs.(14) and (15) in two decoupled equations for sublattices A and B as:

$$(\eta v_F)^2 \left[\begin{array}{l} -\partial_y^2 + \left(\frac{\beta}{a}\right)^2 \varepsilon_{yy}^2 + \left(\frac{eB}{\eta}\right)^2 y^2 + \frac{\beta}{a} (\partial_y \varepsilon_{yy} - \varepsilon_{yy} \partial_y) \\ + \frac{eB}{\eta} (\partial_y y - y \partial_y) + k_x^2 + 2 \frac{\beta}{a} \varepsilon_{yy} k_x + 2 \frac{eB}{\eta} k_x y \\ + 2 \frac{\beta e B}{a \eta} \varepsilon_{yy} y + \frac{1}{4} \left(\frac{g_0 \mu_B B}{\eta v_F}\right)^2 \end{array} \right] \phi_B = \varepsilon^2 \phi_B, \quad (29)$$

$$(\eta v_F)^2 \left[\begin{array}{l} -\partial_y^2 + \left(\frac{\beta}{a}\right)^2 \varepsilon_{yy}^2 + \left(\frac{eB}{\eta}\right)^2 y^2 - \frac{\beta}{a} (\partial_y \varepsilon_{yy} - \varepsilon_{yy} \partial_y) \\ - \frac{eB}{\eta} (\partial_y y - y \partial_y) + k_x^2 + 2 \frac{\beta}{a} \varepsilon_{yy} k_x + 2 \frac{eB}{\eta} k_x y \\ + 2 \frac{\beta e B}{a \eta} \varepsilon_{yy} y + \frac{1}{4} \left(\frac{g_0 \mu_B B}{\eta v_F}\right)^2 \end{array} \right] \phi_A = \varepsilon^2 \phi_A, \quad (30)$$

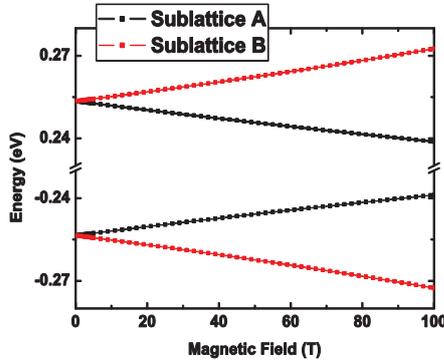


Figure 2: Bandstructures of strained GNRs of sublattices A and B vs magnetic fields. Here we choose the width of the GNR as $L=3\sqrt{3}$ aN ($N=6$). The parameters are chosen as: $\tau_e = -100$ eV/nm, $h_0=1$ nm and $B=1$ T.

RESULTS AND DISCUSSIONS

The schematic diagram of the two-dimensional graphene sheet in a computational domain is shown in Fig. 1 (upper panel). For strained zigzag GNRs, by coupling strain tensor into the Dirac Hamiltonian through pseudomorphic vector potential, we assume that such strain tensor induce a parabolic confinement potential and thus we apply the Dirichlet boundary conditions at the two boundaries to let the wavefunctions to vanish at the boundary and solve numerically of two decoupled Eqs. (29) and (30) numerically based on finite element method (comsol 3.5). For GNRs considered here, typical numbers of elements depend on grid refinements and exceed 800. We solve the multiphysics problem, ensuring the convergence of the results. In Fig. 1(i), we have plotted the wavefunctions squared vs x for the lowest energy states of zigzag GNRs for $k_x=0$. For the case $k_x=0$, the wavefunctions correspond to the nodeless confined states. This is perfectly described by Eqs. (23), (24), (27) and (28). In Fig. 1(ii), we have plotted the band

structures of zigzag GNRs and see that the finite width of the GNR breaks the energy spectrum into an infinite set of bands. Here the solid lines show the confined modes and dotted lines show the edge states of the surface waves that have vanishing energy at or near the edge of the zigzag GNRs. Figs. 1(iii) and (iv) correspond to the wavefunctions and the band diagram of strained zigzag GNRs. Here in Fig. 1(iii) we clearly see that the combination of strain tensor with magnetic fields shift the localization of wavefunctions to the zigzag edge. In Fig. 1(iv) we see that the the strain tensor induce opening of the finite band gap at Dirac point. In Fig. 2, we have plotted energy eigenvalues associated to sublattices A and B of electron hole like states vs magnetic field. Here we see that the spin-splitting energy difference enhances with magnetic field. We re-emphasized the importance of the analysis of coupled effects in such complex physical systems as GNRs.

CONCLUSIONS

Based on analytical and finite element numerical results, we have analyzed the band structures of strained and unstrained zigzag graphene nanoribbons in presence of externally applied magnetic field along z-direction. Our focus has been on coupled effects in such complex systems. In particular, we have shown that finite width of the unstrained GNR breaks the energy spectrum into an infinite set of bands. By implementing the contribution of strain tensor in the Dirac Hamiltonian, we have shown that in the combination of strain tensor and external magnetic field act like a shifted parabolic confinement potential. As a result, we have shown that the localization of wavefunctions of electrons move towards the edge of the zigzag boundary. We have also confirmed that the strain tensor induce opening of the finite band gap at Dirac point. Finally in Fig. 2, we have shown that the Zeeman spin splitting energy enhances with magnetic fields in strained GNRs. The developed model and associated methodology can be useful for the analysis of complex physical systems where coupled physical effects are essential.

ACKNOWLEDGEMENTS

The authors were supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canada Research Chair (CRC) program, and the Bizkaia Talent Grant under the Basque Government through the BERC 2014-2017 program, as well as Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

REFERENCES

- Abanin D. A.; P. A. Lee and L. S. Levitov, 2006. "Spin-Filtered Edge States and Quantum Hall Effect in Graphene.", *Phys. Rev. Lett.* 96, 176803.

- Barbier, M.; P. Vasilopoulos and F. M. Peeters. 2010. "Extra Dirac points in the energy spectrum for superlattices on single-layer graphene.", *Phys. Rev. B* 81, 075438.
- Bao, W.; K. Myhro; Z. Zhao; Z. Chen; W. Jang; L. Jing, F. Miao; H. Zhang; C. Dames and C. N. Lau. 2012. "In Situ Observation of Electrostatic and Thermal Manipulation of Suspended Graphene Membranes." *Nano Letters* 12, 5470.
- Bao, W.; F. Miao; Z. Chen; H. Zhang; W. Jang; C. Dames and C. N. Lau. 2009. "Controlled ripple texturing of suspended graphene and ultrathin graphite membranes.", *Nat Nano* 4, 562.
- Brey, L. and H. A. Fertig. 2006. "Electronic states of graphene nanoribbons studied with the Dirac equation.", *Phys. Rev. B* 73, 235411.
- Castro Neto, A.H.; F. Guinea; N. M. R. Peres; K. S. Novoselov and A. K. Geim. 2009. "The electronic properties of graphene.", *Rev. Mod. Phys.* 81, 109.
- Choi, S.-M; S.-H. Jhi and Y.-W. Son. 2010. "Effects of strain on electronic properties of graphene.", *Phys. Rev. B* 81, 081407.
- Cadelano, E.; P. L. Palla; S. Giordano and L. Colombo. 2009. "Nonlinear Elasticity of Monolayer Graphene.", *Phys. Rev. Lett.* 102, 235502.
- Carpio A. and L. L. Bonilla. 2008. "Periodized discrete elasticity models for defects in graphene.", *Phys. Rev. B* 78, 085406.
- Cadelano, E. and L. Colombo. 2012. "Effect of hydrogen coverage on the Young's modulus of graphene.", *Phys. Rev. B* 85, 245434.
- Cerda, E. and L. Mahadevan. 2003. "Geometry and Physics of Wrinkling.", *Phys. Rev. Lett.* 90, 074302.
- Comsol 3.5a "www.comsol.com."
- Guinea, F.; M. I. Katsnelson and M. A. H. Vozmediano. 2008. "Midgap states and charge inhomogeneities in corrugated graphene." *Phys. Rev. B* 77, 075422.
- Guinea, F.; B. Horovitz and P. Le Doussal. 2008. "Gauge field induced by ripples in graphene.", *Phys. Rev. B* 77, 205421.
- Juan, F. de; J. L. Manes and M. A. H. Vozmediano. 2013. "Gauge fields from strain in graphene.", *Phys. Rev. B* 87, 165131.
- Krueckl, V. and K. Richter. 2012. "Bloch-Zener oscillations in graphene and topological insulators.", *Phys. Rev. B* 85, 115433.
- Kitt, A. L.; V. M. Pereira; A. K. Swan and B. B. Goldberg. 2012. "Lattice corrected strain-induced vector potentials in graphene.", *Phys. Rev. B* 85, 115432.
- Levy, N.; S. A. Burke; K. L. Meaker; M. Panlasigui; A. Zettl; F. Guinea; A. H. C. Neto and M. F. Crommie. 2010. "Strain-Induced PseudoMagnetic Fields Greater Than 300 Tesla in Graphene Nanobubbles.", *Science* 329, 544.
- Landau, L.D. and E. M. Lifshitz. 1970. "Theory of Elasticity (Pergamon Press Ltd)".
- Meng, L.; W.-Y. He; H. Zheng; M. Liu; H. Yan; W. Yan; Z.-D. Chu; K. Bai; R.-F. Dou; Y. Zhang; Z. Liu; J.-C. Nie and L. He. 2013. "Strain-induced one-dimensional Landau level quantization in corrugated graphene.", *Phys. Rev. B* 87, 205405.
- Maksym, P.A. and H. Aoki. 2013. "Magnetic-field-controlled vacuum charge in graphene quantum dots with a mass gap.", *Phys. Rev. B* 88, 081406.
- Novoselov, K.S.; A. K. Geim; S. V. Morozov; D. Jiang; M. I. Katsnelson; I. V. Grigorieva; S. V. Dubonos and A. A. Firsov. 2005. "Two-dimensional gas of massless Dirac fermions in graphene.", *Nature* 438, 197.
- Novoselov, K. S.; D. Jiang; F. Schedin; T. J. Booth; V. V. Khotkevich; S. V. Morozov and A. K. Geim. 2005. "Two-dimensional atomic crystals.", *PNAS* 102, 10451.
- Novoselov, K.S.; A. K. Geim; S. V. Morozov; D. Jiang; Y. Zhang; S. V. Dubonos; I. V. Grigorieva and A. A. Firsov. 2004. "Electric Field Effect in Atomically Thin Carbon Films.", *Science* 306, 666.
- Prabhakar, S.; R. Melnik; L. L. Bonilla and S. Badu. 2014. "Thermoelectromechanical effects in relaxed-shape graphene and band structures of graphene quantum dots.", *Phys. Rev. B* 90, 205418.
- Prabhakar, S. and R. Melnik. 2015. "Relaxation of electronhole spins in strained graphene nanoribbons.", *Journal of Physics: Condensed Matter* 27, 435801.
- Sarma, S.D.; S. Adam; E. H. Hwang; and E. Rossi. 2011. "Electronic transport in two-dimensional graphene.", *Rev. Mod. Phys.* 83, 407.
- Shenoy, V.B.; C. D. Reddy; A. Ramasubramaniam and Y. W. Zhang. 2008. "Edge-Stress-Induced Warping of Graphene Sheets and Nanoribbons.", *Phys. Rev. Lett.* 101, 245501.
- Sevincli, H.; M. Topsakal and S. Ciraci. 2008. "Superlattice structures of graphene-based armchair nanoribbons.", *Phys. Rev. B* 78, 245402.
- Zhang, Y; Y.-W. Tan; H. L. Stormer and P. Kim. 2005. "Experimental observation of the quantum Hall effect and Berry's phase in graphene.", 2005. *Nature* 438, 7065.
- Zheng, H.; Z. F. Wang; T. Luo; Q. W. Shi and J. Chen. 2007. "Analytical study of electronic structure in armchair graphene nanoribbons.", *Phys. Rev. B* 75, 165414.

USING OF ORIENTATION SENSOR CHR6-DM IN SECURITY TECHNOLOGIES

Milan Adámek, Petr Neumann and Martin Pospíšilík
Tomas Bata University in Zlín
Faculty of Applied Informatics
Nad Stráněmi 4511, 760 05, Zlín, Czech Republic
E-mail: adamek@fai.utb.cz

KEYWORDS

CHR6-DM, gyroscope, accelerometer, magnetometer.

ABSTRACT

This article presents the possibilities offered by exploiting CHR 6-DM Orientation Sensors in Security Technologies. It also addresses hardware resources suitable for the inertial navigation of RC models used in Security Technologies. The basis of the CHR 6-DM sensor is a gyroscope that is further complemented by an accelerometer and a magnetometer. In order to use orientation sensor security technology, suitable software was developed that can be used for the autonomous implementation of autonomous RC models suitable for use in Security Technology applications.

INTRODUCTION

Inertial Navigation Systems (further only INS) are used for navigation, i.e. they are mainly used to measure the instantaneous geographic coordinates of a mobile device. The main significance of inertial navigation lies in its autonomy, i.e. its independence from external sources of information. Studies on inertial navigation go back to long before the advent of Micro-Electro-Mechanical System (further only MEMS) technology. The first inertial navigation device was developed and tested by rocket makers like Robert Goddard and Werner Von Braun in the early nineteen-thirties. Later, the inertial technology was further improved by institutions like Drapers Labs - which created the first INS. Inertial navigation enabled further great air accomplishments; e.g. the Apollo rocket and the space program.

Prior to the advent of MEMS technology, precision mechanical gyroscopes and accelerometers had been used. Their high cost however, limited the individual applications to a very great extent.

MEMS

An MEMS micro-system is generally defined as a miniature, intelligent sensing system associating the scanning of information, signal processing and the execution of special functions on the output. Micro-

systems usually combine the properties of two or more of the following six basic energy domains - electrical, mechanical, optical, biochemical, magnetic and thermal. They are usually designed and integrated on a single chip - possibly in a multi-chip hybrid design. Exo-system components have structures with micron (μm) dimensions whose technical features are conditioned by the shape of the microstructure. Micro-systems combine several micro components with two or more functions, optimized in the internal system - in many cases using microelectronic structures and functions.

Prior to the advent of MEMS technology, precise mechanical gyroscopes and accelerometers were used. Further development/evolution of MEMS technology enabled the production of sensors, which - while not attaining the performance of conventional mechanical sensors - but, which dramatically reduced the price, size and weight/mass.

Inertial Measurement Unit (further only IMU/s) micromechanical sensors, have only recently begun to appear on the market cheap, modern and effective alternatives to existing inertial sensors. Many of today's - contemporary, inertial sensors can now be found in military, automobile, industrial, and the pharmaceutical industries - for instance.

INERTIAL UNITS

Inertial navigation represents a navigational technique - where measurement is actualised/implemented using accelerometers and gyroscope. They (are/can be) used for following or tracking positions and for the orientation of objects relative to (their) known starting point, orientation and rapidity.

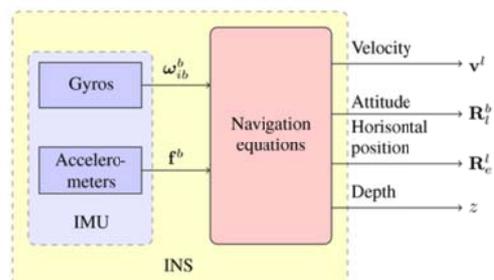


Figure 1: An Inertial Navigation System [1]

An INS typically contains three orthogonally located gyroscopes and three orthogonal accelerometers, which measure declination speed and linear acceleration. The processing of such signals from these devices means the possibility of such devices being used to acquire the location and orientation of objects. Inertial Navigation is based upon the application of Newton's Laws on Motion. The Second Newton Law is used to discover acceleration – according to the following equation:

$$a = \frac{F}{m} \quad (1)$$

where: a is the acceleration of the body, F is the force influencing that body, and m is the body's mass.

A. A Stable Platform System

A Stable Platform System is assembled using inertial sensors, sited upon a platform that is isolated from any form of external rotation.

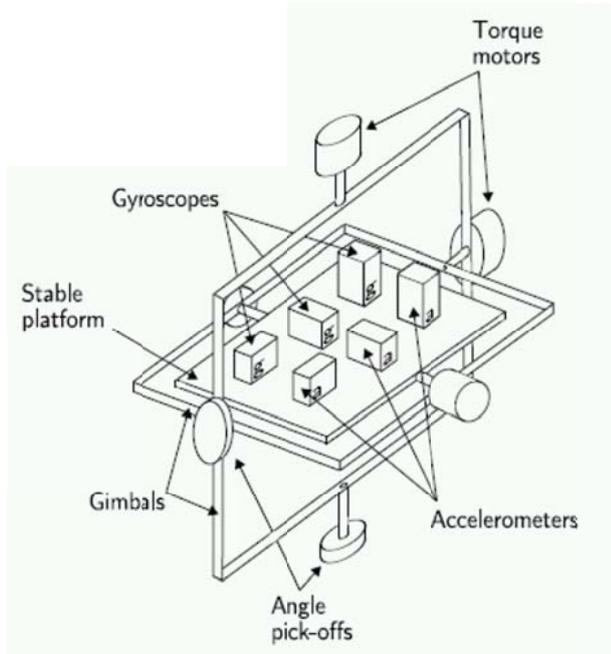


Figure 2: A Stable Platform [2]

B. A “Strapdown” System

In Strapdown systems, inertial sensors are solidly mounted to the device; thus, the output is measured in bodily geometric sets. The integrated output from the gyroscope used for tracking movements is a triad of signals from the accelerometer (which are) transformed into global coordinates. The acceleration signals in global coordinates are then integrated in the same way as in a stable platform algorithm.

Stable platforms and Strapdown systems are based on the same principles. Strapdown systems have reduced mechanical complicatedness – and are physically smaller. These advantages are achieved at the cost of increasing the computational complicatedness, while the

costs for such computational units are reduced. Strapdown systems have become a dominant type of INS.

A TYPICAL SENSOR MODEL

Each MEMS sensor can be described by a linear model.

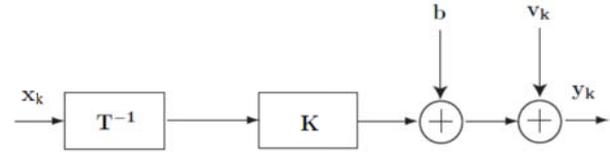


Figure 3: An Inertial Sensor Block Diagram [3]

In Figure 3 is x_k is the value of the physical quantity, T^{-1} is the inverse matrix of the deviated axis, K is the scale, b is bias, v_k is the measurement interference, and y_k is the output of the inertial sensor.

$$b = [b_x \quad b_y \quad b_z]^T \quad (2)$$

$$K = \begin{pmatrix} k_x & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & k_z \end{pmatrix} \quad (3)$$

Since the error of a deviated axis typically attains several tenths of a degree, one can use the following equation:

$$T_a^p = \begin{pmatrix} 1 & -a_{yz} & -a_{zy} \\ a_{xz} & 1 & -a_{zx} \\ -a_{xy} & a_{yx} & 1 \end{pmatrix} \quad (4)$$

The measured output can then be modelled as follows:

$$y_k = K(T_a^p)^{-1} x_k + b + v_k \quad (5)$$

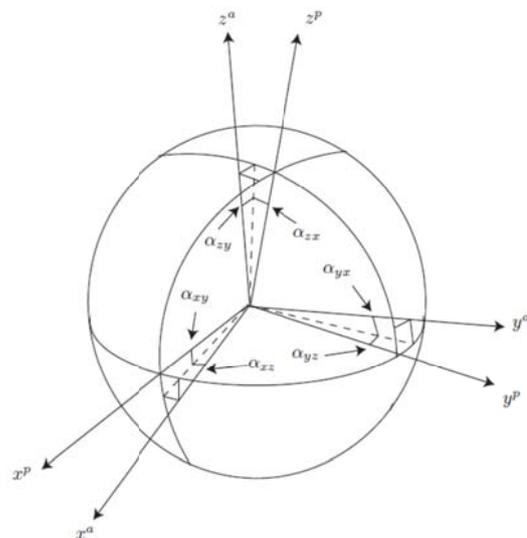


Figure 4: An Axis Deviation Error [3]

INERTIAL SENSOR QUALITY

All types of accelerometers and gyroscopes demonstrate bias, scale errors, errors in the inclination of the scale axis, or random interference. The size of these errors depends upon the type of sensor.

Every source of systematic errors has four components: fixed contributions, temperature deviations, run-to-run deviations, and in-run deviations. Fixed contributions are permanently present on a sensor and the INS processor unit is corrected with the aid of data measured in a laboratory. As regards Run-to-run, the error size changes over time on the sensor – but remains constant for any run. The in-run contribution deviations change slowly in the course of the activity. Theoretically, it is possible to correct this error through the addition of more sensors; but, practically speaking, this is very difficult to attain.

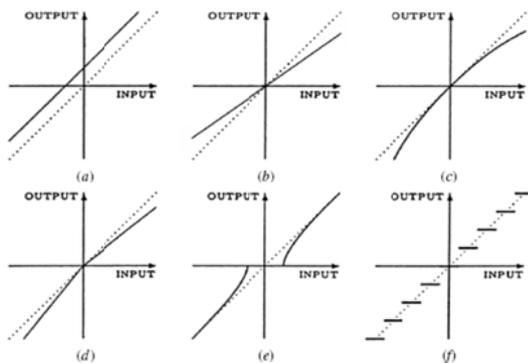


Figure 5: Typical Inertial Navigation Errors: (a) Bias, (b) Scale Errors, (c) Non-linearity Errors, (d) Asymmetrical Errors, (e) Dead-zone Errors, (g) Quantificational Errors [4]

A. Bias

Bias is a constant error that occurs in all accelerometers and gyroscopes. In many cases, bias is a dominant error in inertial sensors.

Table 1. Bias Error Levels for IMU Degrees [5]

IMU Grade	Accelerometer Bias		Gyro Bias	
	mg	$m s^{-2}$	$^{\circ} hr^{-1}$	$rad s^{-1}$
Marine	0.01	10^{-4}	0.001	5×10^{-9}
Aviation	0.03-0.1	$3 \times 10^{-4} - 10^{-3}$	0.01	5×10^{-8}
Intermediate	0.1-1	$10^{-3} - 10^{-2}$	0.1	5×10^{-7}
Tactical	1-10	$10^{-2} - 10^{-1}$	1-100	$5 \times 10^{-6} - 5 \times 10^{-4}$
Automotive	>10	$>10^{-1}$	>100	$>5 \times 10^{-4}$

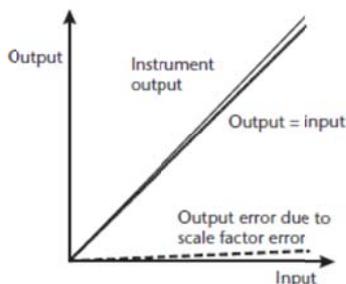


Figure 6: Scale Error [6]

B. Scale Errors

A scale error is a deviation in the input-output inclination. Accelerometer output error is dependent upon the size of the acceleration force acting upon the axis. For gyroscopes – this is on the size of angle velocity.

C. Deviated Axis Errors

Deviated axis errors occur in all types of INS. This error is the consequence of technological restrictions on production.

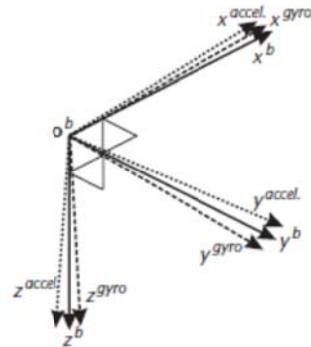


Figure 7: Deviated Axes of Accelerometers and Gyroscopes [6]

D. Non-linearity

Non-linearity represents the precision of real calibration curves with the ideal static transmission characteristics (i.e. straight line). It expresses in percent the upper border of the scale's extent and provides the maximum deviation of any calibration point whatsoever from the corresponding point on the ideal characteristic. Sensor linearity error (i.e. Non-linearity), is defined as follows:

$$L_e = \frac{\Delta y_{\max}}{\text{Full scale}} \quad (6)$$

INERTIAL NAVIGATION HARDWARE RESOURCES

A. STEVAL-MKI02V2

The STEVAL-MKI02V2 kit is a development kit from STMicroelectronics.

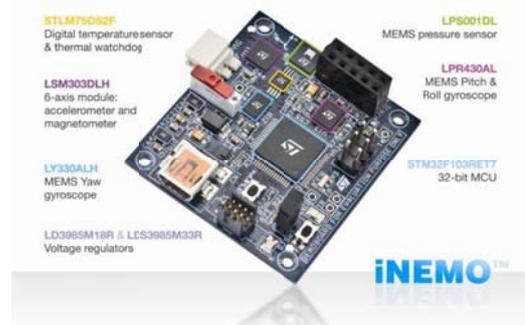
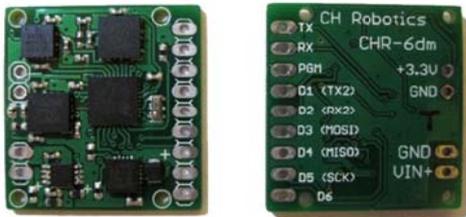


Figure 8: The STEVAL_MKI02V2 Development Kit [7]

- Two possible ways of charging: A charging connector or a USB connector
- STM32F103RE: High-density performance line, ARM-based 32-bit MCU with 256 to 512 kB Flash, USB, CAM, 11 timers, 3 ADC and 13 communication interfaces
- LPR430AL: 2-axis gyroscope (roll, pitch), 300°/s full extent with an analogue output and settable additional filters
- LSM303DLH[7]: 6-axis geomagnetic module: $\pm 2 \text{ g} / \pm 4 \text{ g} / \pm 8 \text{ g}$ acceleration range, configurable magnetic field $\pm 1,2$ do 8,1 Gauss (max), I2C Bus-box
- LPS001DL: Pressure sensor range: 300-1100 mbar with an I2C Bus-box
- STLM75: Temperature sensor range: -55 to +125°C and an I2C Bus-box

B. CH Robotics 6DM

The CH Robotics CHR-6DM AHRS contains an IMU complemented by a triple-axis magnetometer. Data from all of the sensors are collated by a 64 MHz ARM Cortex M3 processor with an Expanded Kalman Filter (EKF). The EKF combines data from the accelerometer, gyroscope and magnetometer with estimates of yaw, pitch, and roll angles. The output is presented in Euler



Degrees via a UART Bus-box with speeds up to 300 Hz.

Figure 9: CH Robotics CHR-6DM AHRS Developer Kit [8]

Main components:

- STMicroelectronics LPR510AL – pitch a roll gyroscope: $\pm 100^\circ/\text{C}$, with analogue output
- STMicroelectronics LY510LH – yaw gyroscope
- Analogue Device: ADXL335[8] – tri-axial accelerometer
- Honeywell HMC5843 – tri-axial digital magnetic compass

Functions:

- EKF estimation of yaw, pitch and roll angles
- Adjustable output speeds: (20 Hz – 300 Hz)
- Motherboard with a 3.3V regulator
- +3.3 output, with an ability of up to 400mA for recharging other peripherals (e.g. GPS)
- Two UARTs and an SPI Bus-box

A SOFTWARE DESIGN FOR THE CHR6-DM SENSOR

In order to be able to track all of the measured quantities from the sensors on-line, a user-program was written using the C# language. The program communicates with the measurement system across a UART or RS232 interface.

Communication between the measurement system and the user-program takes place across a UART interface with a transmission speed of 115200, 8 bits, without parity, with a single stop bit. For data transmission purposes, a data packet was created whose length is 41 bits.

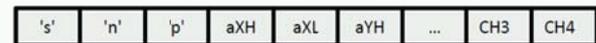


Figure 10: Data Transmission Packet for the User-program

The meaning of the individual bytes is as follows:

- 1 to 3 – bits: 's','n','p' – form the headliner of the message
- 4 – bit: aXH [mg] – is the upper byte of acceleration on axis X, written in paired additions
- 5 – bit: aXL [mg] – is the lower byte of acceleration on axis X, written in paired additions
- 6 to 7 – bits: aYH, aYL [mg] – is the acceleration along axis Y, also written in paired additions
- 8 to 9 – bits: aZH, aZL [mg]
- 10 to 15 – bits: gXH, gXL, gYH, gYL, gZH, gZL [°/s] – are the angle speeds in individual axes, measured by the gyroscopic sensor
- 16 to 21– bits: mXH, mXL, mYH, mYL, mZH, mZL [mGauss] – are the outputs from the magnetometer in the individual axes
- 22 to 27 – bits: rollH, rollL, pitchH, pitchL, yawH, yawL [°/s] – are the rotational angles of the inertial unit

- 28 to 41 – bits: rollRateH, rollRateL, pitchRateH, pitchRateL, yawRateH and yawRateL – are the angle speed, acquired by merging the gyroscopic sensor, the accelerometer and the magnetometer.

The following figure represents an example of the depiction of the data from the gyroscope, accelerometer and magnetometer in the user-environment.

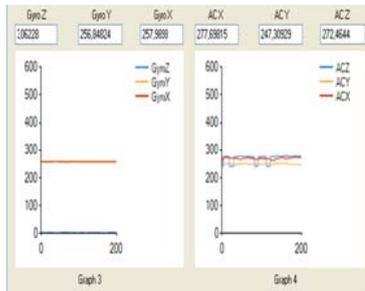


Figure 11: Depiction of the course of the output data from the gyroscope and accelerometer

The CHR6-DM sensor was used for the analysis of the behaviour of the individual control components of the RC model. To be able to track on-line all of the measured quantities on the RC model, software was designed using C# language which communicates with the measurement system across the UART interface.

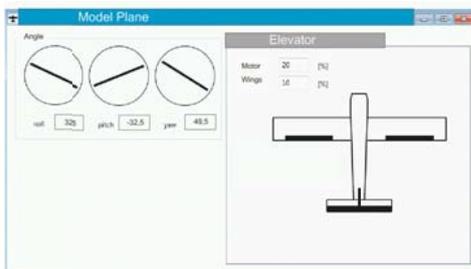


Figure 12: User-program for the depiction of the measured quantities of an RC model

The entry entitled “Acceleration” depicts the information about all of the measured quantities on the CHR 6-CD kit. The program also enables the record data to files function. Data recording is possible in two ways. The first is the recording of all of the data received by the program. The first button “Record Data” registers all the data. The second possibility is to select only a certain section of the data – where the second button “Allow Selection” serves for this function. Upon clicking this button, an index is saved of the last-received data and a second click finishes the “data collection” while the recording of this data is written in a file.

CONCLUSION

The system suggested herein is a suitable instrument for the control of pilotless RC aircraft, which can be used in security applications. Especially, it enables the exploitation of such RC models for monitoring hard-to-access locations during floods - or, as the case may be during extensive fire outbreaks. It is becoming clear that it is appropriate to construct such pilotless RC models with inertial navigation capacities.

Contemporary technical levels already allow the development of very high-quality IRS (Inertial Reference Systems) with precise measurement of positional angles in flight of cca $\pm 1^\circ$.

The precision and reliability of an IRS conceived in this way can be substantially improved – for instance, by the use of a minimum of three groups of inertial sensors (very cheap, light, small, and with minimum consumption).

In the case where stochastic processes breakdowns causing measurement errors are correlated, and after adding together the adequate signals, the resultant measurement errors are reduced to one third. An IRS like this, with micromechanical (MEMS) inertial sensors’ precision approaches that of an IRS, artificial horizon and gyro-magnetic compass based on classical gyroscopic technology and, in the coming decade, it can be expected that this will completely force out classical gyroscopic technologies from onboard aircraft.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014).

REFERENCES

- [1] Fauske, K. M. *Inertial navigation system* [Online]. 2006. Available from: <http://www.texample.net/tikz/examples/inertial-navigation-system/>
- [2] Woodman, O. J. *An introduction to inertial navigation* [Online]. 2007 Available from: <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-696.pdf>.
- [3] Naranjo, N. *Analysis and Modelling of MEMS based Inertial Sensors*. Stockholm, 2008; Kungliga Tekniska högskolan, School of Electrical Engineering. Available from: https://eeweb01.ee.kth.se/upload/publications/reports/2008/XR-EE-SB_2008_011.pdf
- [4] Grewal, S. Weill, R. Angus P. *INERTIAL SYSTEMS TECHNOLOGIES: Systematic Errors*. [Online] Available from: <http://beta.globalspec.com/reference/14795/160210/chapter-9-3-2-inertial-systems-technologies-systematic-errors>
- [5] Vectornav. *High Quality Orientation Sensors*. [Online]. [cit. 18.5.2012]. Available from: http://www.vectornav.com/index.php?option=com_content&view=article&id=12&Itemid=15#KalmanFilter

- [6] Groves, P. *Principles of GNSS, Inertial, and Multi-Sensor Integrated Navigation Systems* (GNSS Technology and Applications). 2008. BOSTON|LONDON: ARTECH HOUSE. ISBN-13: 978-1-58053-255-6
- [7] STMicroelectronics. *User manual*. Version 1. STEVAL-MKI062V2. [Online]. Available from: http://www.st.com/internet/com/TECHNICAL_RESOURCES/TECHNICAL_LITERATURE/USER_MANUAL/CD00271225.pdf
- [8] CH Robotics. *Datasheet. CHR-6dm*. [Online]. Available from: http://www.chrobotics.com/docs/chr6dm_datasheet.pdf
- [9] Hong, Park, Sungsu, S. K. 2008. *Minimal-Drift Heading Measurement using a MEMS Gyro for Indoor Mobile Robots*. Open Access Sensors, Vol. 8, p. 1 - 13.

Informatics of the Tomas Bata University in Zlín, Czech Republic. His current research covers the following topics: electromagnetic compatibility, shielding effectiveness of materials for avionics, design of construction of electrical circuits and testing of electrical devices considering the security of communication. The security issues are investigated in cooperation with Escola Superior de Tecnologia e Gestão, Beja, Portugal. His e-mail address is: pospisilik@fai.utb.cz.

AUTHOR BIOGRAPHIES



MILAN ADÁMEK graduated in 1990 from the Olomouc Palacky University, Czech Republic. He received his Ph.D. degree in Technical Cybernetics at Tomas Bata University in Zlin in 2002. From

1997 to 2008 he worked as senior lecturer at the Faculty of Technology, Brno University of Technology. From 2008 he has been working as an associate professor at the Department of Electronic and Measurement, Faculty of Applied Informatics of the Tomas Bata University in Zlín, Czech Republic. Current work covers following areas: power lines, camera system, sensors. His e-mail address is: adamek@fai.utb.cz.



Petr NEUMANN has been graduated from the Brno Technical University in Electronic Technology in 1974. He has acquired the industrial experience in the field of medical

electronics and quality management as R&D engineer. He received his Ph.D. degree in Technical Cybernetics at Tomas Bata University in Zlin in 2001. He has been lecturing and working in the university research area since 1994. He was engaged in the SMT technology training, equipment installation and servicing more than 10 years between 1997 and 2009. He is currently working as a senior lecturer at Tomas Bata University in Zlin. His research work is aimed at the electronic component authenticity analysis and failure diagnosis. His e-mail address is: neumann@fai.utb.cz



MARTIN POSPÍŠILÍK graduated in 2008 from Czech Technical University in Prague, Czech Republic, in Microelectronics. Having received his Ph.D. degree in Engineering Informatics

at Tomas Bata University in 2013, he became an assistant and researcher at the Department of Computer and Communication Systems of Faculty of Applied

SCHOTTKY DIODE REPLACEMENT BY TRANSISTORS: SIMULATION AND MEASURED RESULTS

Martin Pospisilik
Department of Computer and Communication Systems
Faculty of Applied Informatics
Tomas Bata University in Zlin
760 05, Zlin, Czech Republic
E-mail: pospisilik@fai.utb.cz

KEYWORDS

Model, SPICE, MOSFET, Electronic Diode, Voltage Drop, Power Dissipation

ABSTRACT

The software support for simulation of electrical circuits has been developed for more than sixty years. Currently, the standard tools for simulation of analogous circuits are the simulators based on the open source package Simulation Program with Integrated Circuit Emphasis generally known as SPICE (Biolk 2003). There are many different applications that provide graphical interface and extended functionalities on the basis of SPICE or, at least, using SPICE models of electronic devices. The author of this paper performed a simulation of a circuit that acts as an electronic diode in Multisim and provides a comparison of the simulation results with the results obtained from measurements on the real circuit.

INTRODUCTION

For the first time, Simulation Program with Integrated Circuit Emphasis (SPICE) was released in 1973 as a general-purpose, open source analog electronic circuit simulator. It was intended to check the integrity of circuit designs on the board-level and to predict the circuit behaviour in time (Vladimirescu 1994). In 2011 the development of SPICE has been named and IEEE Milestone (Bogdanowicz 2011).

To process the simulation, the devices of the simulated circuit must be defined in the form of the set of parameters. These parameters can be set manually, or, obtained from the manufacturers of the devices. The simulation itself is based on numerical methods and its complexity exceeds the framework of this paper.

The author has chosen the application Multisim, released by National Instruments, that employs SPICE models and algorithms, but provides graphical interface and a wide variety of device libraries. The aim was to provide a comparison of results obtained by means of simulation with the results measured on the real circuit.

MOTIVATION

The expansion of metal oxide field effect transistors has led to efforts to substitute conventional diodes by electronic circuit that behave in the same way as the diodes, but show considerably lower power dissipation as the voltage drop over the transistor can be eliminated to very low levels. In Fig. 1 there is a block diagram of a backup power source for the devices that use Power over Ethernet technology. This solution, that has been developed at Tomas Bata University in Zlin, enables immediate switching from the main supply to the backup one together with gentle charging of the accumulator, unlike the conventional on-line uninterruptable power sources do (Pospisilik and Neumann 2013).

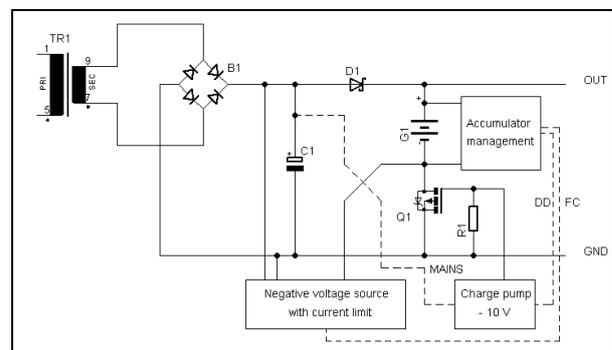


Figure 1: Simplified Block Diagram of the Backup Power Source for Power over Ethernet Devices

When the power supply network fails, the Charge pump is activated. This results in opening of the Q1 transistor. Now the output is fed by the accumulator and the Negative voltage source is decommissioned, being blocked by the Schottky diode D1. Once the supply network starts to be active again, the Charge pump is switched off and the transistor Q1 is closed as its gate charge is distracted by the R1 resistor. The Negative voltage source is fed from the rectifier B1, delivering controlled charging voltage drop at the currents around 6 A is expected to be as high as 1 to 1.2 V, resulting in the power dissipation of up to 7.2 W. Therefore there was a need to find a simple and stable solution that would be implementable on a small area of the printed circuit board, ideally by means of surface mounted

devices, not needing any other heatsink than the copper on the board. As the solution, a functional sample of a circuit employing a transistor instead of the diode has been created and its behaviour was simulated and tested.

The principle of replacement of the Schottky diode with the metal oxide field effect transistor (MOSFET) is depicted in Fig. 2. If the P-channel MOSFET is used, the current flows in the direction from the drain (D) to the source (S) of the transistor. At the voltages lower than the threshold voltage between the gate (G) and the source of the transistor (S) the current flows through the internal protective diode. When the voltage is increased and the voltage difference between the source and the gate of the transistor is higher than its threshold voltage, the transistor is turned to the ON state and its conductivity is increased significantly. Once the polarity of the power source is alternated, the transistor does not lead any current at all. However, the maximum voltage difference between the gate and the source of the transistor is limited and therefore in practical solutions the circuit cannot be as simple as depicted in Fig. 2.

Thanks to the fact that there is the Negative voltage source equipped with a current limiter connected in series with the accumulator, the voltage and current stresses to the accumulator are considerably limited. The bottleneck of this solution is the Schottky diode D1. Its expected

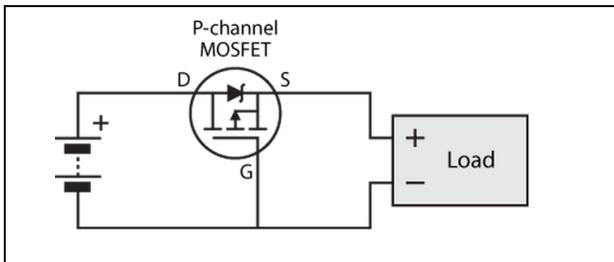


Figure 2: Replacement of a Diode by P-channel MOSFET

CIRCUIT DESIGN

The diagram of the circuit that can replace the Schottky diode D1 depicted in Fig. 1 can be found in Fig. 3. The principle of the operation of the circuit is as follows: When there is a power source connected to the pads P1 and P2 and the voltage on P2 is higher than the voltage on P1, the current starts to flow through the internal diode of the transistor Q1 to the load that is connected between pads P3 and P4. The voltage drop at this diode and at the diode D2 is high enough to open the transistor T2 that drives the transistor T1. As the transistor T1 starts to conduct the current, a voltage drop is created on the resistor R1, resulting in generation of the sufficient gate-to-source voltage on the transistor Q1. At this moment, the transistor Q1 is opened, exhibiting a good conductivity. Now the voltage drop over the transistor is very low which leads to pinching of transistor T2 and the transistor T1 respectively. This effect helps to keep

the gate-to-source voltage of the transistor T1 at the appropriate level at various input voltages.

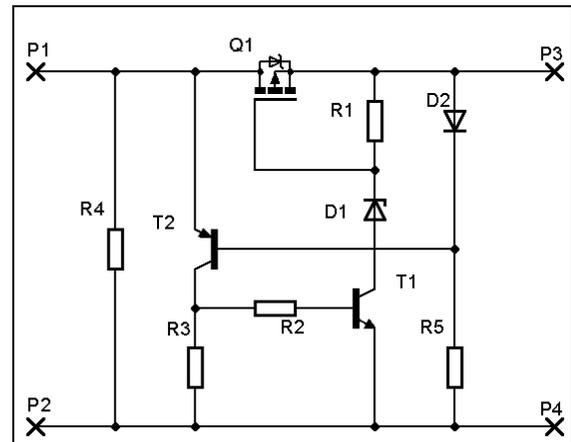


Figure 3: The Circuit Replacing the Schottky Diode D1 in Fig. 1

The description of the devices used in the circuit is provided in Table 1.

Table 1: Elements of the circuit depicted in Fig. 3

Element	Description
R1	100 k Ω , metal oxide resistor
R2	1 k Ω , metal oxide resistor
R3	47 k Ω , metal oxide resistor
R4	330 k Ω , metal oxide resistor
R5	10 k Ω , metal oxide resistor
D1	Zener diode 12 V, BZX55C12
D2	Diode, 1N4001
T1	NPN transistor, BC547
T2	PNP transistor, BC557
Q1	MOSFET, IRF9Z34N

SIMULATION

The simulation was performed in Multisim 12 according to the model depicted in Fig. 4. As can be seen, the values of resistors have been set according to Table 1 and their tolerances have been set to 1 %. The semiconductors D1, D2, Q2 and Q3 have been chosen from the software's library. The model of the transistor Q1 was implemented according to the information provided by its producer. This model is shown in Table 2.

For the purposes of simulation, the circuit was loaded with an adjustable current source I1 whereas the source of energy for the circuit was simulated as a pre-settable piecewise linear voltage source V1.

Different kinds of simulation of the circuit has been performed as described in the subchapters below.

Table 2: Spice model of the transistor Q1 (Vishay 2010)

```

*Mar 30, 2010 *Doc. ID: 90320, Rev. A
*File Name: part irf9z34n_sihf9z34n_PS.txt and part
irf9z34n_sihf9z34n_PS.spi
*This document is intended as a SPICE modeling guideline and does
*not constitute a commercial product datasheet. Designers should
*refer to the appropriate data sheet of the same number for
* guaranteed specification limits.
.SUBCKT irf9z34n 1 2 3
*****
* Model Generated by MODPEX *
*Copyright(c) Symmetry Design Systems*
* All Rights Reserved *
* UNPUBLISHED LICENSED SOFTWARE *
* Contains Proprietary Information *
* Which is The Property of *
* SYMMETRY OR ITS LICENSORS *
*Commercial Use or Resale Restricted *
* by Symmetry License Agreement *
*****
* Model generated on Apr 12, 99
* MODEL FORMAT: SPICE3
* Symmetry POWER MOS Model (Version 1.0)
* External Node Designations
* Node 1 -> Drain
* Node 2 -> Gate
* Node 3 -> Source
M1 9 7 8 MML=100u W=100u
* Default values used in MM:
* The voltage-dependent capacitances are
* not included. Other default values are:
* RS=0 RD=0 LD=0 CBD=0 CBS=0 CGBO=0
.MODEL MM PMOS LEVEL=1 IS=1e-32
+VTO=-3.18176 LAMBDA=0 KP=2.52466
+CGSO=4.9266e-06 CGDO=1e-11
RS 8 3 0.0001
D1 1 3 MD
.MODEL MD D IS=2.51148e-12 RS=0.0124373 N=1.05244 BV=55
+IBV=0.00025 EG=1 XTI=2.91741 TT=0.0001
+CJO=4.87958e-10 VJ=5 M=0.731488 FC=0.5
RDS 3 1 1e+06
RD 9 1 0.028942
RG 2 7 6
D2 5 4 MD1
* Default values used in MD1:
* RS=0 EG=1.11 XTI=3.0 TT=0
* BV=infinite IBV=1mA
.MODEL MD1 D IS=1e-32 N=50
+CJO=8.50824e-10 VJ=0.5 M=0.456256 FC=1e-08
D3 5 0 MD2
* Default values used in MD2:
* EG=1.11 XTI=3.0 TT=0 CJO=0
* BV=infinite IBV=1mA
.MODEL MD2 D IS=1e-10 N=0.4 RS=3e-06
RL 5 10 1
FI2 7 9 VF12 -1
VF12 4 0 0
EV16 10 0 9 7 1
CAP 11 10 8.50824e-10
FI1 7 9 VF11 -1
VF11 11 6 0
RCAP 6 10 1
D4 6 0 MD3
* Default values used in MD3:
* EG=1.11 XTI=3.0 TT=0 CJO=0
* RS=0 BV=infinite IBV=1mA
.MODEL MD3 D IS=1e-10 N=0.4
.ENDS irf9z34n

```

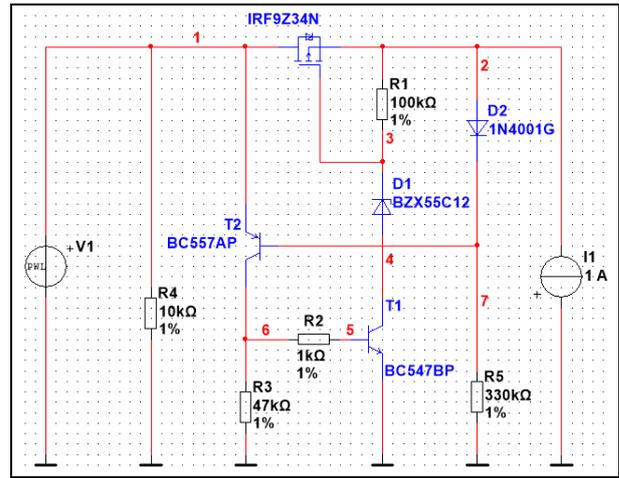


Figure 4: Simulation Schematics of the Circuit Created in Multisim 12

Voltage Drop at Different Loads

In the next step, the voltage drop between the input and the output of the circuit has been simulated for different output loads and different input voltage. For this purpose, the PWL source (V1) was set to increase its voltage from 0 to 30 V in the period from 0 to 1 s and transient analysis of the circuit was performed repeatedly for different settings of the current source I1.

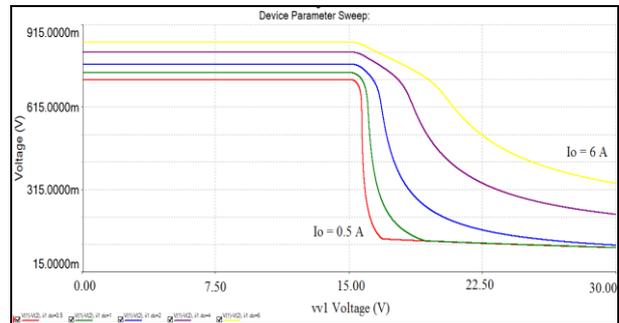


Figure 6: Voltage Drop over the Circuit Versus the Load Current

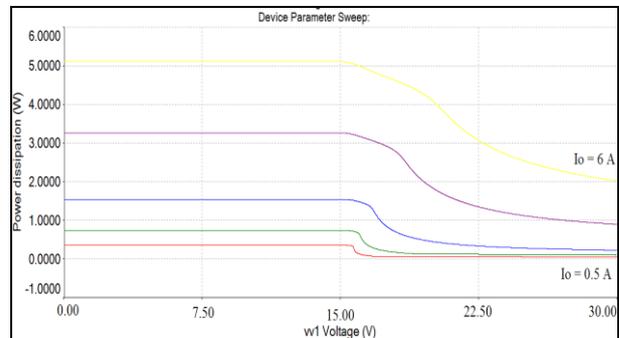


Figure 7: Calculated Power Dissipation on the Transistor Q1

The Parameter sweep simulation has been used to perform this in one step. In the graph that is depicted in Fig. 6 there are the output results for the load currents 0.5, 1, 2, 4 and 6 A. The output results consist in the difference between the input and the output voltage of the circuit (e.g. the voltage drop).

The power dissipation on the transistor Q1 can also be obtained directly by the simulation as depicted in Fig. 7.

Rectification at 50 Hz

The behavior of the circuit used as a rectifier operating at standard 50 Hz frequency has always been simulated. The modification of the simulation schematics was done as depicted in Fig. 8. The power source was replaced by a standard AC power source with the output voltage of 20 V_{RMS} and the frequency of 50 Hz. The circuit was loaded by a combination of a 4 Ω resistor and 47 mF capacitor in parallel. The internal impedance of the AC power source was set to 30 mΩ.

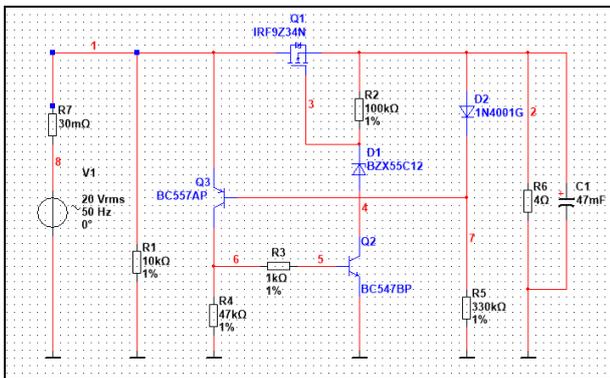


Figure 8: Simulation Schema for the Circuit Used as a Rectifier

This situation simulates the operation of a one-way rectifier connected to a transformer with a low output impedance. The waveforms simulated in the nodes 1 and 2 can be found in Fig. 9.

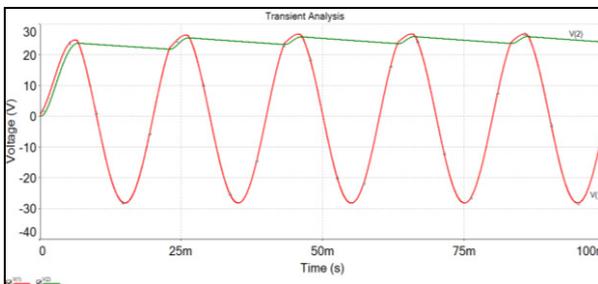


Figure 9: Waveforms at the Input and the Output of the Circuit When It Is Used as a 50 Hz Rectifier

RESULTS OBTAINED BY MEASUREMENT

In order to prove the results obtained by simulation, a functional sample of the circuit has been constructed and its behavior has been tested by means of the following laboratory equipment:

- Programmable power source Picotest P9611A,
- Electronic load Array 3721,
- Digital oscilloscope Hameg HMO722,
- RMS Multimeter UNI-T UT803,
- Multimeter Voltcraft VC820.

The configuration of the experiment was made according to the scheme depicted in Fig. 10.

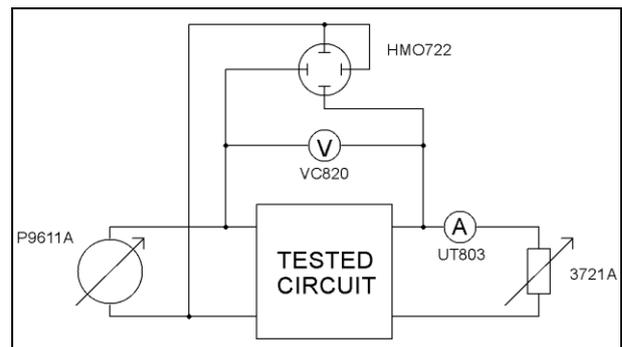


Figure 10: Experiment Setup

Considering the capabilities of the equipment, only DC characteristics of the electronic diode circuit have been measured.

From the measurements that have been obtained, measurement of the voltage drop over the circuit at different loads was the most interesting one. The voltage of the power source has been increased from 0 to 25 V with the step of 1 V and the electronic load was sequentially set to a constant current of 0.5, 1, 2, 4 and 6 A. The voltage drop over the circuit was measured by the voltmeter VC820 and the stability of the DC voltage at the input as well as at the output of the circuit was monitored by the oscilloscope HMO722. The output current was monitored by the multimeter UT803.

The results of the measurement can be found in Fig. 11.

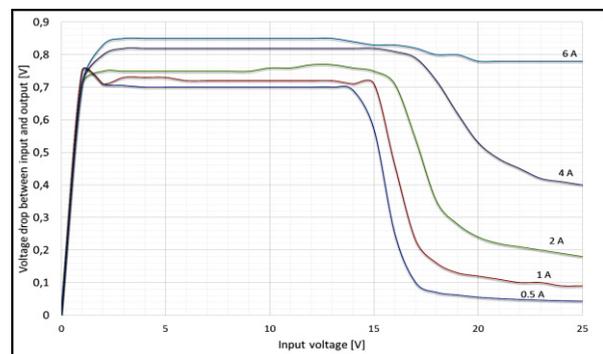


Figure 11: Measured Voltage Drop Over the Circuit at Different Current Loads

The findings resulting from the experimental measurement can be compared to the results of simulation of the same operating conditions of the circuit that are depicted in Fig. 6. For low currents the measured results correspond to the results obtained by the simulation. However, at higher currents the results obtained by the simulation were not correct. It is expected, that the increased voltage drop, obtained by the real experiment, has been partially caused by the resistance of the connecting wires as well as by the resistance of the PCB traces. At the current of 6 A and the input voltage of 22.5 V, considering the voltage drop over the current, the difference between the simulation and the measurement was as high as 0.3 V, which corresponds to the traces' resistance of about 50 m Ω .

CONCLUSIONS

In this paper the construction of the circuit that replaces a conventional Schottky diode is described together with the results of simulations that were made in order to verify the circuit's design before it was constructed as well as with the results obtained by measurement on the functional sample of this circuit. It was proven that the circuit operates correctly and the achieved results were close to the simulated behavior of the circuit with the exception of high current loads. When the circuit was loaded with high currents, the voltage drop over it was considerably higher than simulated.

Acknowledgements

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the

National Sustainability Programme project No. LO1303 (MSMT-7778/2014).

REFERENCES

- Biolek, D. 2003. *Solving of electrical circuits [Resime elektronické obvody]*. BEN – Technická literatura. ISBN 80-7300-125-X.
- Vladimirescu, A. 1994. *The SPICE book*. John Wiley & Sons. ISBN 978-0471609261.
- Bogdanowicz, A. 2011. "SPICE Circuit Simulator Named IEEE Milestone". *The institute*. IEEE.
- Pospisilik, M., Neumann, P. 2013. "Improved Design of the Uninterruptable Power Supply Unit for Powering of Network Devices". In *19th IEEE International Symposium on Design and Diagnostics of Electrical Circuits*. Karlovy Vary, Czech Republic. ISBN 978-1-4673-6133-0.
- Vishay. 2010. *IRF9Z34 SPICE model*. Online: <http://www.vishay.com/docs/90320/sihf9z34.lib>

AUTHOR BIOGRAPHY



MARTIN POSPISILIK was born in Přílepy, Czech Republic. He reached his master degree at the Czech Technical University in Prague in the field of Microelectronics in 2008. Since 2013, after finishing his Ph.D. work focused on a construction of the Autonomous monitoring system, he became an assistant professor at the Tomas Bata University in Zlin, focused on communication systems and electromagnetic compatibility of electronic components. His e-mail is: pospisilik@fai.utb.cz

Simulation and Optimization

A NEW APPROACH FOR THE BULLWHIP EFFECT

Hans-Peter Barbey

University of Applied Sciences Bielefeld
Interaktion 1, 33619 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 orders, material flow and stock is rather difficult. In order to recognize the basic relationships between the parameters, a very simple model was set up. It consists of 4 identical stages. In all stages the stock is closed-loop controlled to a nominal stock. Therefore the only decision which can be done in the entire supply chain is the quantity of an order. In a first simulation run a suitable order strategy will be defined. Good results can be realized, if an order is splitted up in two: A customers order and a stock order. In a second run this strategy will be applied to a seasonal trend of the customers requirements. It will be shown that the bullwhip effect can be minimized with the applied order strategy.

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. A not defined 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 that 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 minimize 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.

A very simple model of a supply chain without any cooperation between the particular members 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 controlling strategies have been applied to a seasonal trend in this simple simulation model. (Barbey 2014). Now this model will be used with a controlling strategy, which includes a kind of cooperation between the members of the supply chain. The model is designed in the following manner:

The model consists of four identical stages according fig. 1. The behavior of each stage is the same. The time to place an order is 1 time unit (TU). The time for delivery is 3 time units. Therefore lead time to fill up the stock for one stage is the sum of both, 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 from the very beginning to the end of the supply chain. To be able to fulfill a customers order within the minimum lead time of 4 TU each stage needs a stock.

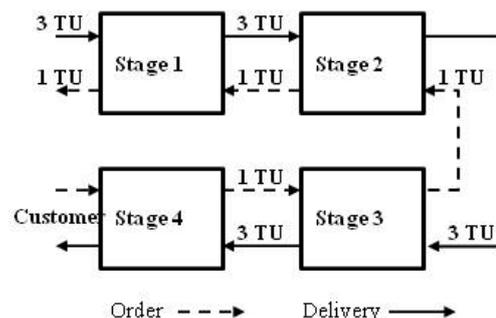


Figure 1: Model of a Supply chain
(TU= time unit)

The only decision, which can be done in this simulation, is to decide about the quantity of the order. This order has two tasks: It fulfills the predecessors order in the supply chain and compensates a difference in the own inventory. 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 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 a more impressive manner. To demonstrate the bullwhip effect clearly, all other influences like delay in delivery or empty stock have been eliminated.

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 the simulation model as described above.

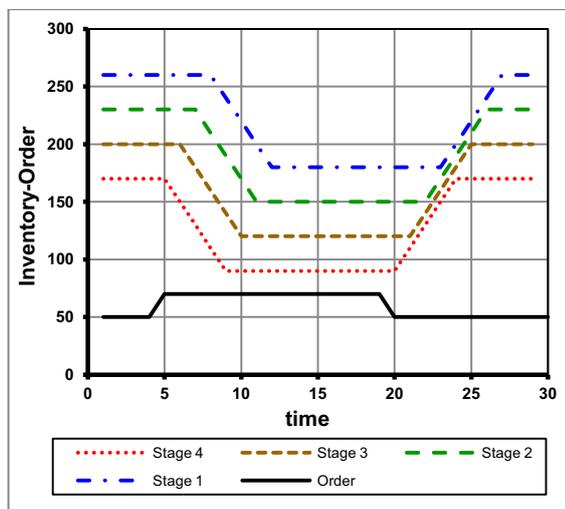


Figure 2: Stock with input=output strategy

If the customer increase his order, here from 50 to 70 items, the stock of stage 4 decrease in a linear manner (fig. 2). 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.

An improved strategy is a one-order-strategy. That means a stage orders material at his supplier, which covers the requirements of his customer and compensates the deviation in his own stock. Assuming the increase or decrease in the order is permanent, the aim of each particular stage is to equalize this difference, which occurred with the strategy “input is output”, to the nominal stock. Therefore the orders have to be increased for a certain time above the customer order (fig. 3). In this example the time for compensation is 16 time units in one particular stage. 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. Only the stage at the very end of

the supply chain (stage 4) is able to compensate the stock difference within the scheduled time, here 16 time units (fig.3 and fig. 4). For all other stages it requires more than double the time.

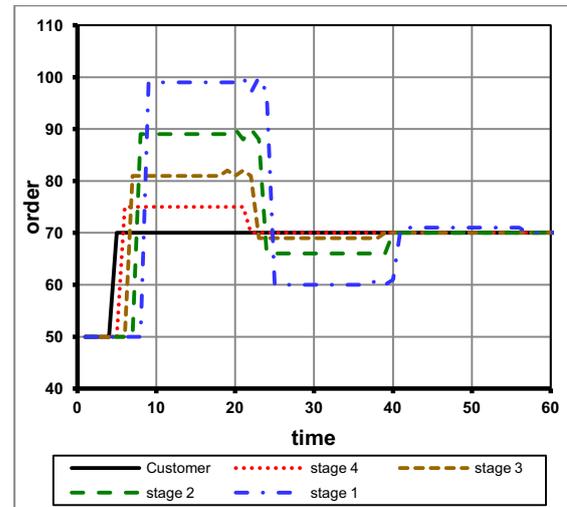


Figure 3: One-orders-strategy with compensation within 16 time units

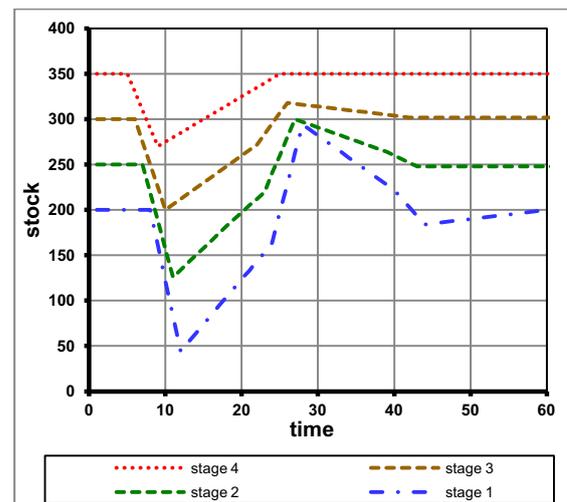


Figure 4: Stock with one-order-strategy: constant order within 16 time units

This is quite obvious: The last stage has only to fulfill the customers requirement. All other stages have to fulfill the customers 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).

The second strategy seems to be relative similar, but it is quite different. The best strategy to fulfill a customers order is:

$$\text{Order in} = \text{order out}$$

This strategy leads to a deviation of the stock from the nominal stock in each stage of the supply chain as explained above. To fill up the stock to the nominal stock has nothing to do with a customers order, it is only

related to the behavior of a particular stage of the supply chain. Therefore a second order, the stock order, should be done. The only decision is now how long the compensation of the stock will take.

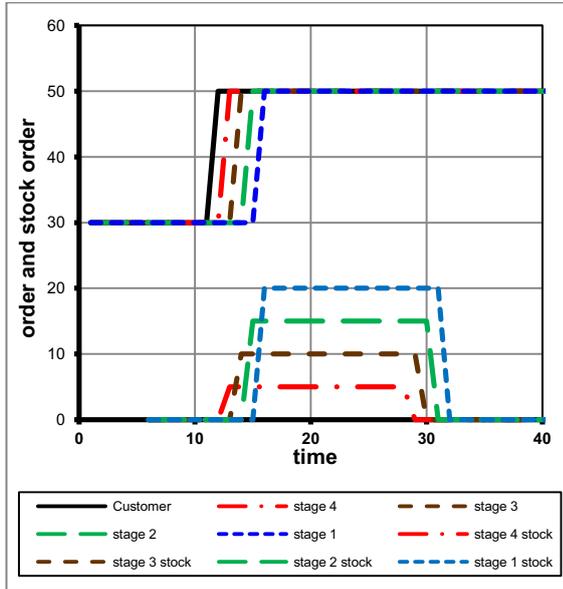


Figure 5: Order strategy of the closed-loop controller: Customers order and stock order with a compensation within 16 time units

Fig. 5 shows the in=out strategy for the customers order and the stock order with a compensation time of 16 to eliminate the stock deviation. The customers order is the same for all stages only with a time difference of one time unit. The stock order increases from stage to stage. This is obvious because a stage has to compensate his own stock difference and all

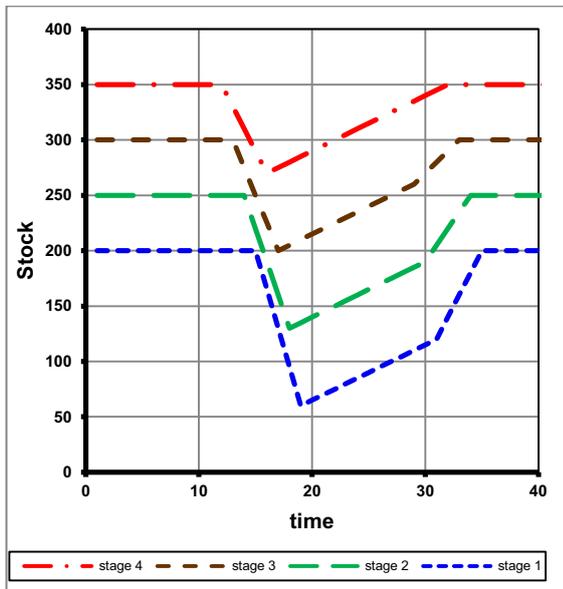


Figure 6: Stock compensation to the nominal stock within 16 time units

differences of the stages downstream. Therefore the bullwhip effect is only created by the stock orders. Each stage upstream has a higher stock difference (fig. 6). The bullwhip effect occurs in the stock difference

too. However each stage can compensate the stock difference in the same time.

Comparing both strategies for the 2nd one there are three advantages:

- All stages can compensate their stock differences in the same time
- The total order (customer order + stock order) is lower
- The bullwhip effect in the stock difference is slightly lower

This strategy is perfect, if customers order changes only one time. If there is a linear trend in the orders of the customer (fig. 7), a permanent deviation in the stock occurs (fig. 8).

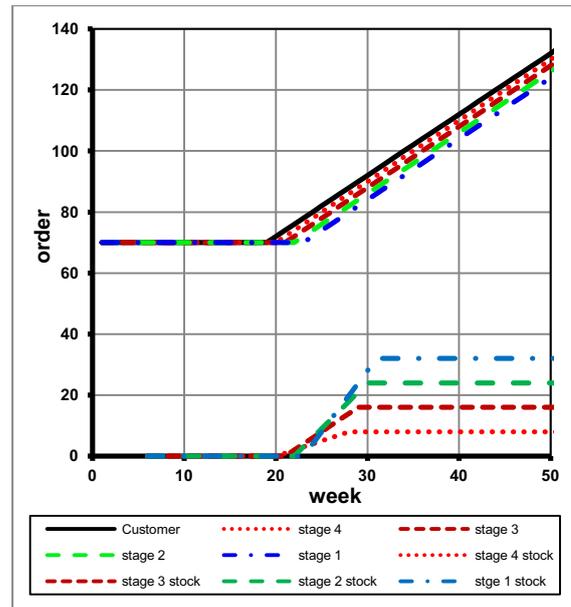


Figure 7: Customer order and stock order for a linear trend with a compensation time of 8 TU

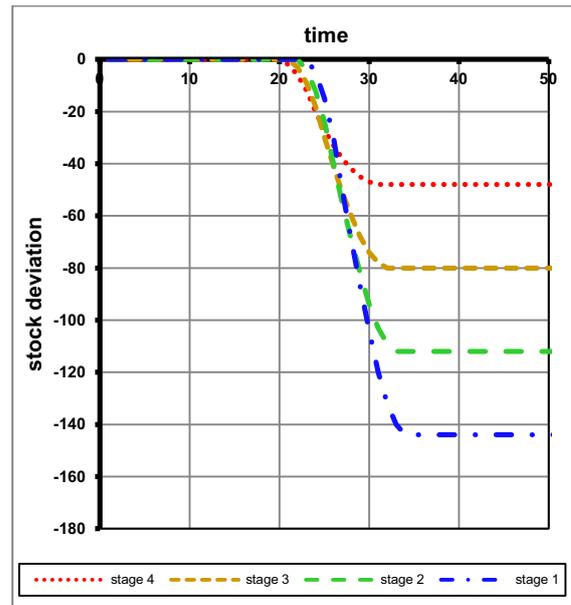


Figure 8: Permanent stock deviation caused by a linear trend in the orders, compensation time 8

When the linear trend starts the stock orders have a linear increase. After the compensation time they come in a steady state (fig. 7) and the stock deviation is in a steady state too (fig. 8). It can be shown that this permanent deviation depends from the increase of the trend, the duration of compensation time for the stock order and the position of a particular stage in a supply chain. In the next step the deviation has been calculated with these parameters and was included in the stock order (fig. 9). After starting the trend there is an increase of the stock orders over a time length of the compensation time. After that time the stock orders are constant with the same values as in fig. 7.

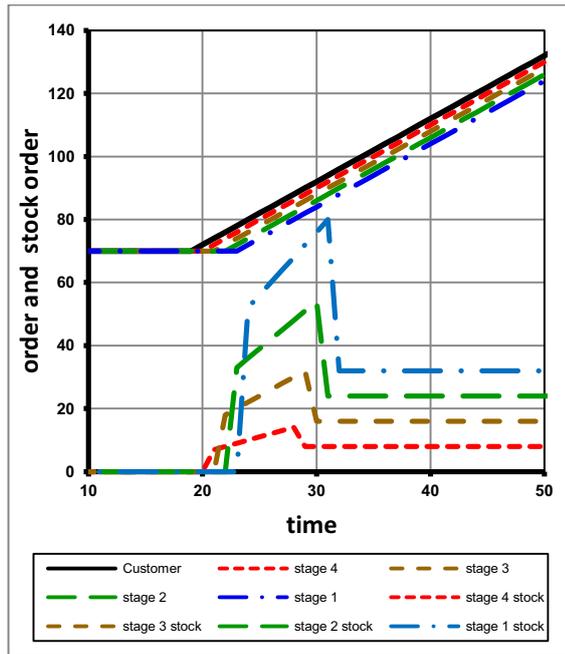


Figure 9: Compensation of the linear trend with a stock order, compensation time 8

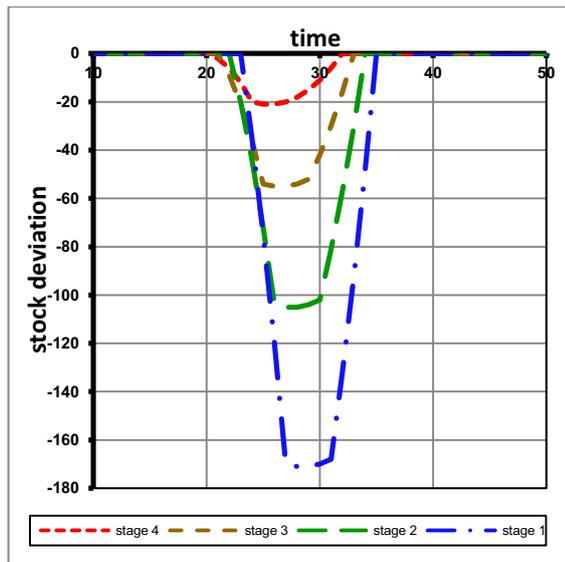


Figure 10: Stock deviation with the compensation of the linear trend, compensation time 8

For the first three stages the deviation from the nominal stock is better than without the calculation of this

compensation. For the last stage the deviation is worse (fig. 10). But all stages can reduce the deviation to zero. A further examination of the linear trend will be not done in this paper. Seasonal trends seem to be more important.

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 an 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 +/- 40% of the average, which is 500 in this simulation. The period of this sine is 300 time units. The following simulations examine the fluctuation of the stock for the individual stages in the supply chain and the variations in the orders. Three different control strategies are applied:

1. Order in = order out
2. One-order-strategy (fig. 11 and fig. 12)
3. Customer order and stock order including compensation of a trend. (fig. 13 and fig. 14)

The first strategy is not a real controlling strategy. It is only applied to get a basis to compare the other strategies. The variation in the orders according to the sine from minimum to maximum is 400. In all stocks the variation of the stock items from minimum to maximum is 1600 (fig.11).

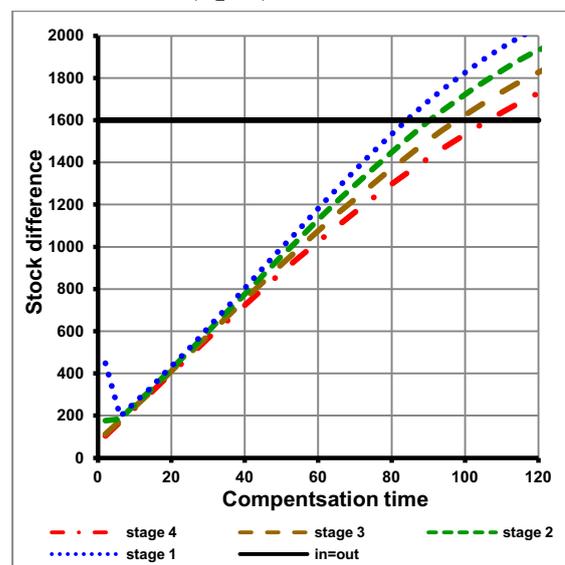


Figure 11: Stock difference with strategy 2

Important for the other strategies of the closed-loop control is the duration of the compensation time. Therefore in the next simulation runs varies the compensation time from 2 to 120.

For strategy 2 exists for very short compensation times a bullwhip effect in the stock. Then the stock difference diminishes to a minimum and increases again with elongation of the compensation time (fig.11). At a compensation time of 80 for stage 1 and 100 for stage for the stock difference becomes worse than with the in=out strategy. Now the closed-loop controller is too slow to compensate the variation in the stocks. For the order differences occurs an extreme bullwhip effect especially for stage 1 (fig.12). After a minimum the order differences increases slowly again with an increasing compensation time. It is obvious that in=out strategy has better results all the time. The reason is that with this strategy no additional stock order for compensation has to be created.

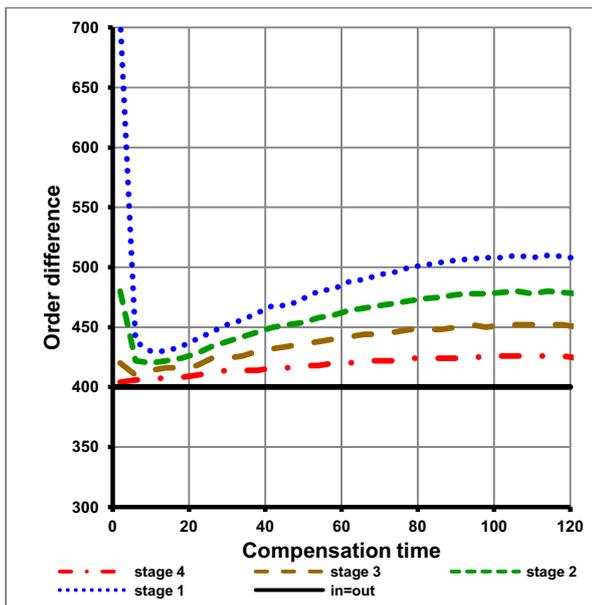


Figure 12: Order difference with strategy 2

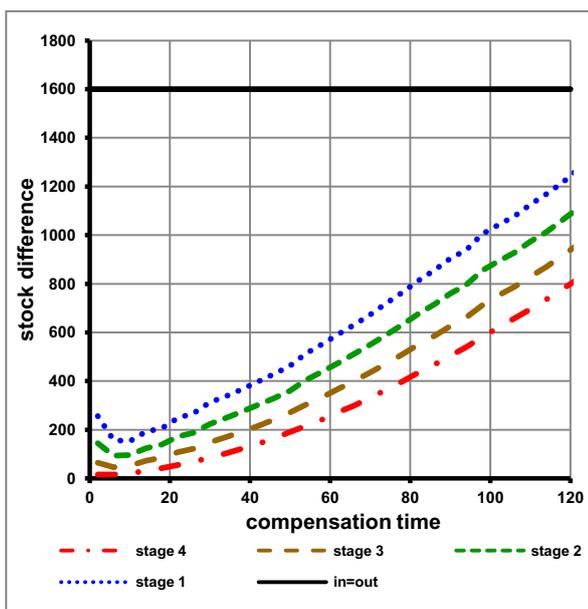


Figure 13: Stock difference with strategy 3

Much better results can be realized with strategy 3 (fig.13). Just as with strategy 2 a bullwhip effect exists a short compensation times too. After a minimum the stock differences increase with an increasing compensation time. However, even with large compensation times the results are much better than with the in=out strategy. The order differences are very similar to strategy 2 (fig.14). Only some rounding effects caused by the simulation occur in the diagram.

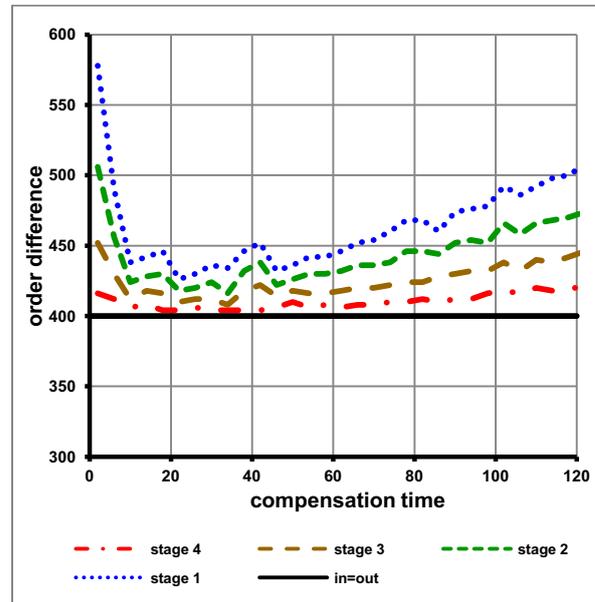


Figure 14: Order difference with strategy 3

CONCLUSIONS AND SUMMARY

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 customers orders. This has been realized by a closed-loop control. In this 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 strong influence on the stock. Two effects can minimize the stock. First it should be applied a short compensation time. Is that time too short, a bullwhip effect can occur. Second a split of the order should be done: Customers order and stock order. The customers is handled like the in=out strategy and only the stock order is close-loop controlled. This split of the order is a kind of cooperation between the members of a supply chain: A supplier of a stage in the supply chain gets information in terms of the order about the customer of that stage.

REFERENCES

- Barbey, H.-P.: Seasonal Trends in Supply Chains. Proceedings of 28. European Conference on Modelling and Simulation (ECMS), Brescia, 2014, 748-752.
- 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 BUSINESS PROCESS SIMULATION STANDARD (BPSIM): CHANCES AND LIMITS

Ralf Laue
Department of Computer Science
University of Applied Sciences of Zwickau
Dr.-Friedrichs-Ring 2a, 08056 Zwickau, Germany
Ralf.Laue@fh-zwickau.de

Christian Müller
Faculty of Business, Computing, Law
Technical University of Applied Sciences Wildau
Hochschulring 1, D-15745 Wildau, Germany
christian.mueller@th-wildau.de

KEYWORDS

Event driven simulation, business processes, process analysis, Business Process Modeling and Notation (BPMN), Business Process Simulation Interchange Standard (BPSim)

ABSTRACT

This paper provides a critical analysis of the BPSim standard, a specification by the Workflow Management Coalition. The aim of this standard is to make it possible to exchange simulation models between different modeling and simulation tools. We discuss the expressiveness of BPSim model and come to the conclusion that it will be sufficient for certain cases, but also lacks some important features.

INTRODUCTION

Business processes models are usually specified in graphical languages. The most popular standard for such a language is Business Process Model and Notation (BPMN) [BPMN 2013].

For simulation purposes, the models have to be enriched by additional information and transformed into formal specifications that can be processed by a simulation tool [Anthony Wallner et. al. 2006, Raimar Scherer 2011]. The Business Process Simulation Interchange Standard (BPSim) is a BPMN extension for process simulation. It was developed by some industrial actors (Fig. 1) and published as a standard specification [BPSim 2013, BPSim 2014].

Not all products of the contributors are fully supporting the BPSim specification. Known implementations were provided by Trisotech, Lanner, Sparx and jBPM.

A BPSim simulation engine is not only an extension of an BPMN engine, because the aim of a BPMN engine is process automation and not simulation. Such an automation engine must store its data persistently in a database. Simulation runs must be fast, hence a simulation engine should store the data in internal memory. For this reason, implementing a simulation engine in a BPMN suite requires considerable effort.

In this paper, we discuss the main ideas of BPSim as well as its chances and limits.



Figure 1 BPSim Contributors

BPMN DIAGRAMS AND SERIALIZATION

Before we start discussing BPSim as an extension of BPMN2 (version 2.0 is the current version of the BPMN standard), we will describe the basic ideas of BPMN2. Its aim is the modeling of business processes for documentation and automation purposes. The standard [BPMN] defines the graphical representation of models, its semantics and an XML-based serialization format. In Fig. 2, we show a BPMN diagram that we will use as an example throughout this paper. First, a decision task is executed. At a subsequent gateway, the process path splits depending on the outcome of the decision.

The basic (simplified) XML file structure for this process fragment is shown in Fig. 3. A definitions element contains the required resources and processes.

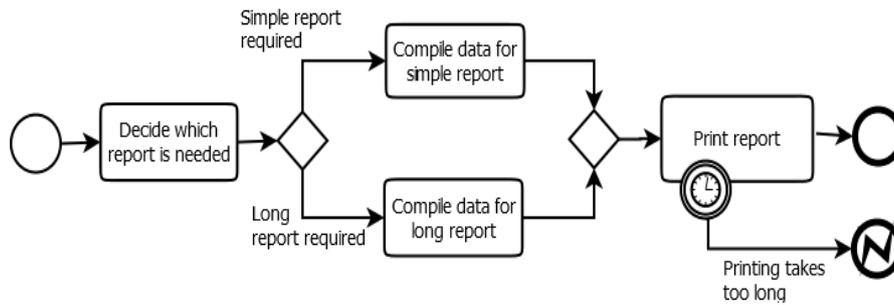


Figure 2 Example Process

The *process* element includes its events, tasks, gateways and sequence flow arcs. The XML file contains also information about resources and data that is not shown in the diagram: The *resource* tag describes the performer of a task and has a model-wide scope. In contrast to this, the *property* element belongs to a certain process. It describes a variable with a scope of a process instance. It is also possible to use a so-called datastore for variables with a model-wide scope (not just referring to a single process execution). In our example, we did not use this feature.

The decision logic is modeled at the outgoing arcs of the gateway by 2 expressions, deciding which arc is used.

```

<definitions xmlns="...">
  <resource id="resource:id" />
  <process id="process_id" >
    <property id="property_id" name="report"/>
    <startEvent id="start_id" />
    <sequenceFlow sourceRef="start_id" targetRef="task_id"/>
    <task id="task_id">
      <performer id="performer_id" >
        <resourceRef>resource_id</resourceRef>
      </performer>
    </task>
    <sequenceFlow sourceRef="task_id" targetRef="gatw_id"/>
    <exclusiveGateway id="gatw_id" />
    <sequenceFlow sourceRef="gatw_id" targetRef="task2_id">
      <conditionExpression>
        <![CDATA[decision == 'simple_report'];]>
      </conditionExpression>
    </sequenceFlow>
    <sequenceFlow sourceRef="gatw_id" targetRef="task3id">
      <conditionExpression>
        <![CDATA[decision == 'long_report'];]>
      </conditionExpression>
    </sequenceFlow>
    ....
  </process>
  <bpmndi:BPMNDiagram><!--Refers to graphical diagram layout--></bpmndi:BPMNDiagram>
</definitions>

```

Figure 3

BPSIM EXTENSION

BPMN supports extensions for different purposes. One such extension is the Business Process simulation

Interchange Standard (BPSim) [BPSim 2013, BPSim 2014] for simulating business processes. Using the *bpsim_namespace* (<http://www.bpsim.org/schemas/1.0>), it adds additional information to an XML serialization in the BPSimData node (Fig. 4).

A BPSim model is organized in different simulation scenarios with start, duration, seed and replication values (Fig 5). A parameter for a warm-up period is missing, but some vendors have extended their implementation by this parameter. For working with variations of scenarios, the inheritance of scenarios is supported.

```

<definitions>
  <resource /> <process /> ..
  <bpmndi:BPMNDiagram> </bpmndi:BPMNDiagram>
  <relationship type="BPSimData">
    <extensionElements>
      <bpsim:BPSimData xmlns:bpsim="
        "http://www.bpsim.org/schemas/1.0">
        </bpsim:BPSimData>
      </extensionElements>
    </relationship>
  </definitions>

```

Figure 4

It is possible to assign simulation parameters to BPMN nodes and arcs by means of an *elementParameter* in a scenario definition. E.g., it is possible to define a *durationParameter* for task nodes (Fig. 5).

Various time parameters (such as setup time or processing time) can be added to tasks. Control parameters allow to define how often / with which probability certain events or certain decisions occur. Resources can be defined and it is possible to assign resources to tasks. In addition, tasks can have priorities. Fixed costs and costs per unit can be assigned to both tasks and resources. All these attributes can be defined depending on calendar definitions, for example by specifying that the availability of resources depend on workdays or shifts.

All the parameters described above can be either fixed values, historical data series or defined as realization of random variables that follow a certain distribution. The standard allows to use 13 types of distributions (that can be parametrized) for this purpose. This is clearly

positive – the authors are aware of several business process simulation tools that work with a too limited set of distributions.

```
<bpsim:Scenario id="default" name="Scenario" ....>
  <bpsim:ScenarioParameters
    replication="2" seed="999" ... >
    <bpsim:Start>
      <bpsim:DateTimeParameter value="2016-01-01T00:00:00"/>
    </bpsim:Start>
  </bpsim:ScenarioParameters>
  <bpsim:ElementParameters elementRef="task_id">
    <bpsim:TimeParameters>
      <bpsim:ProcessingTime>
        <bpsim:DurationParameter value="PT1H"/>
      </bpsim:ProcessingTime>
    </bpsim:TimeParameters>
  </bpsim:ElementParameters>
</bpsim:Scenario>
```

Figure 5

In a BPMN model, properties and data objects are used to control the execution of the model instances. In addition, BPSim allows to add properties to each node and arc by means of *ElementParameters*. The relation between BPMN and BPSim parameters is not specified and depends on the simulation engine, e.g., it is not clear how a model should be interpreted if its BPMN2 parameters for resource usage contradict to the BPSim *ResourceParameters*.

For extending our model to a classical simulation model with a capacity of e.g. 10 resources and a requirement of e.g. 2 resources per task we can use the following parameters of Fig. 6 (all belonging to the BPSim namespace):

```
<ElementParameters elementRef="resource_id">
  ...<ResourceParameters>
  .....<Quantity>
  .....<NumericParameter value="10"/>
  .....</Quantity>
  ...</ResourceParameters>
</ElementParameters>

<ElementParameters elementRef="task_id">
  ...<ResourceParameters>
  .....<Selection>
  .....<ExpressionParameter value=
    "bpsim:getResource('resource_id', 2)" />
  .....</Selection>
  ...</ResourceParameters>
</ElementParameters>
```

Figure 6

This definition (as given in [BPSim 2014]), runs in the simulation engine from Lanner, but it contradicts to the standard specification ([BPSim 2013, Sect. 7.3]) that defines that *ResourceParameters* are not associated to a *task_id* but to the *performer_id*, which belongs to the task element.

For modeling a time schedule, the standard allows calendar-depended parameters. Unfortunately, this feature is not supported by all tools.

Parameters can be marked as *ResultRequest* (Fig. 7) in order to collect the results of a simulation run. This allows to ask for minimum, maximum and mean values (for example of costs or durations), sums (aggregated values, e.g., total time spent in a certain task) and for the number of occurrences (for example of a certain event).

```
<ElementParameters elementRef="task_id">
  ...<TimeParameters>
  .....<WaitTime>
  .....<ResultRequest>sum</ResultRequest>
  .....</WaitTime>
  .....<ProcessingTime>
  .....<ResultRequest>sum</ResultRequest>
  .....</ProcessingTime>
  ...</TimeParameters>
</ElementParameters>
```

Figure 7

CHANCES AND LIMITS OF BPSIM

A lot of simulation models for business processes have simple scenarios. For such cases, the BPSim approach (adding parameters to BPMN elements by means of the BPSim extension) works well: For these models its a great improvement that BPSim allows formulating simulation models independently from modeling tools and simulation engines.

However, in a practical test, we found that on the one hand, some tools implement only a subset of the standard. On the other hand, they provide useful (but proprietary) vendor extensions. E.g. the Trisotech modeler does not support the assignment of a resource parameter to a task. For running a simulation with resources, the model must be changed by hand in text editor. Hopefully, such problems will be solved by time.

A reason for the current situation may be that BPSim is a new standard and there currently only a few competitors on the market.

However, the BPSim specification has also some structural problems that will be discussed in the following sections.

Use of Expression Parameters has Limits

For many scenarios, adding parameters to BPMN elements by means of the BPSim extension works well. For more complicated cases, BPSim allows to add properties to a process instance as well as to BPMN elements.

For example, in our reporting process, a property of a process instance could be the number of pages of the report. Such process instance properties can be regarded

as global variables that can be read and written in the context of each BPMN element. If the upper path in the diagram of Fig. 1 is taken, a simple report has to be compiled (number of pages = 30) while otherwise a long report has to be compiled (number of pages = 100). When the decision has been taken, the property “numberOfPages” is set to the appropriate number. Using so-called expression parameters, it is also possible to define that the costs for the task “print report” depends on the number of pages (say 2 cents per page). This way, by reading and modifying properties, some additional logic (that cannot be seen in the BPMN diagram) can be added to the model.

```
<bpsim:CostParameters>
...<bpsim:fixedCost>
.....<bpsim:ExpressionParameter value=
    "bpsim:getProperty('numberOfPages') *0.02"/>
</bpsim:fixedCost>
</bpsim:CostParameters>
```

Figure 8

In this case, BPSim specifies that the parameter *bpsim:CostParameters* (Fig. 8) is serialized as an XML element, and the content of this element can be an expression parameter.

The situation is different if we try to model the processing time of the print task in the same manner. We can specify that the time is given by a truncated normal distribution with a mean of 70 (and a minimum and maximum value) as follows (Fig. 9):

```
<bpsim:TimeParameters>
<bpsim:ProcessingTime>
<bpsim:TruncatedNormalDistribution
    max="1000" mean="70" min="0"
    standardDeviation="10"/>
</bpsim:ProcessingTime>
</bpsim:TimeParameters>
```

Figure 9

In Fig. 9, the distribution parameters such as “mean” are *attributes* (in this case with the data type Double) in the XML serialization, and the standard provides no means to express them as a calculated value (i.e. as an BPSim expression parameter) as it was the case for the costs.

BPSim Semantics is Interrelated with BPMN Semantics

Let’s assume that we want to interrupt the task “print report” if it took more than 5 minutes. This can be expressed in plain BPMN: A boundary timer event is added to the task (see Fig 2), and it is provided with a *TimerEventDefinition* attribute that specifies that the event fires 5 minutes after the task has been started (Fig. 10):

```
<boundaryEvent id="cancelPrintTimer_id"
    name="Cancel Print" cancelActivity="true"
    attachedToRef="PrintTaskID">
  <timerEventDefinition>
    <timeDuration>PT5M</timeDuration>
  </timerEventDefinition>
</boundaryEvent>
```

Figure 10

In this case, the timing behavior is completely defined in BPMN (not using the BPSim extension), and BPSim does not provide a standard way to say “interrupt the task if it took more than 5 seconds multiplied with the current value of the “number of pages” attribute. Although the timing behavior of the boundary event could be defined using BPSim as well, this would not be appropriate because the BPSim attribute *InterTrigger Timer* that would have to be used for this purpose cannot be related to the point of time when the task “print report” has been started. What would be needed, but is not included in the standard, is the possibility to deal with different timers which can be reset when a task starts (or in general: when an event occurs).

Resource Model not yet Fully Elaborated

Next, let’s assume that before starting to print, the printer needs a warm-up period of 3 minutes if the last print job ended more than 20 minutes ago. For modeling such a situation, the possibility to reset a timer (this time when a task ends) would be required again. In addition, it would be useful if the printer (a resource) would have a time parameter denoting the needed warm-up time as well as a property parameter for storing the information when the last print job ended. Unfortunately, according to BPSim, both kinds of parameters are not allowed for resources.

Altogether, BPSim uses an advanced, but not yet fully elaborated resource model. Resources can have more than one role. Priorities can be assigned to activities. Also, the availability of resources can be defined depending on time intervals (e.g. representing shifts). It is possible to model the (un)availability of resources as a random variable in order to deal with illness or malfunction of technical resources. However, there are still things missing. Although activities can have priorities, the standard does not say anything about the semantics of such a priority attribute. It can be assumed that the meaning of priorities is that a resource when becoming available is assigned to the activity with the highest priority (a feature that [Wal06] requires for useful business process simulation) – but it should be possible to define other resource allocation strategies as well. Even if two activities have no priority information (or both have the same), it should be possible to specify whether a resource is allocated to a random activity, to the most recent one (LIFO), to the one which has been waiting for the longest time (FIFO), etc.

There is no means for modeling consumable resources (such as raw material) that will not be released when a

task that needs a resource is completed. While such information could be modeled as a property of a process instance, such a way of modeling is less intuitive than having a richer resource model. In particular, resources should be allowed to have user-defined property parameters (which is currently not the case).

A richer resource model would also be very useful for modeling working preferences, locations and working speed of resources. In [Wil M. P. van der Aalst et. al 2009], it is discussed that current simulation tools often use oversimplified resource models. Among others, it is not taken into account that people do not work on a constant speed and tend to work part times or in batches. Other than assuming that a resource is available as soon as it is required by a task, simulation models should be able to support various resource patterns [Nick Russel et. al. 2005]. While the support of such rich resource models has been announced as one of the goals of the BPSim initiative [Jan11], the resource model in version 1.0 of the standard has still room for improvement.

Working with historical process data sets

Often, simulation models have a lot of parameters such as duration times, interarrival times and probabilities for decisions. Accordingly, a lot of replications are required to get statistical valid interpretations of the simulation.

In a typical process improvement projects, the data of historical process instances are known from the logs of BPM engines. In order to build a realistic simulation model, it makes sense to use randomly generated values only for those parts of the model, for which no historical data are available. This approach reduces the number of randomly generated parameters, and the number of required replications can be reduced considerably.

BPSim supports working with historical datasets. In a BPSim model, these datasets are assigned to simulation parameters such as decisions or duration times of tasks. However a weakness of the approach is that this assignment is always done in the context of the whole process and does not refer to process instances. In our example (Fig. 1), this would mean that historical data can be used for simulating the decisions and the duration of the tasks in the process. However, the fact that the task “print report” takes longer when the decision “long report required” has been taken, would not be considered in the model.

Result Types are Insufficient

A weakness of the BPSim standard is that the result types that can be requested from a simulation are too limited. Allowed result types are the number of occurrences, minimum, maximum and mean values. However, average, best and worst case scenarios (represented by the minimum and maximum values) are often not enough to describe the statistical distribution

of the simulation results. At least, an information about standard deviation and skewness is desirable. In addition, we have to ask for percentiles as well if we want to deal with service-level agreements such as “95% of the requests have to be handled within n time units”. Unfortunately, such descriptors have not been considered in the BPSim standard. In general, it would be desirable to require that a simulation tool should write a log (in a standardized format) containing all events and decisions happening during a simulation run. This would allow any analysis after a simulation run.

CONCLUSIONS

The motivation behind the BPSim specification was to close the gap between the well-established BPMN standard for modeling and a great variety of simulation tools, each one requiring a proprietary input format. Having such a standard can help to promote the use of business process simulation and to build tools for modeling simulation models independently from simulation tools.

BPMN and BPSim are powerful enough to create models for business process which are „static“ in the sense that parameters may be random variables, but the distribution of those random variables does not change during the process. However, we see from the above examples that the BPSim standard needs improvement for cases where probability distributions change when the process is executed.

Also, it has to be noted that neither BPMN nor BPSim has a fully elaborated resource model. For simulation purposes, a more detailed metamodel for resources (a suggestion can be found in [Cristina Cabanillas 2011]) would be desirable.

Additionally, the semantics of BPMN and its extension BPSim can lead to contradictions. Also, historical data and result types do not support the im- and export of raw data. At this point BPSim should be extended.

Neither in the investigation for [Christian Müller et. al. 2015a and 2015b] nor in the preparation on this paper we found tools that have a full BPSim support. All current tools support the standard partially and have additional vendor extensions. In one case it was necessary to modify a model generated by a BPSim modeler with a text editor for running it in a simulation engine. These examples show that, in contradiction to the BPMN environment, the interchangeability of models between tools is not yet satisfactory. The authors hope that this will be changed by time.

REFERENCES

- Wil M. P. van der Aalst; Joyce Nakatumba; Anne Rozinat and Nick Russell 2009: Business Process Simulation: How to get it right? International Handbook on Business Process Management, Springer, 2009

- BPMN 2013: ISO/IEC International Standard 19510: Information Technology – Object Management Group Business Process Model and Notation, Document Number ISO/IEC 19510:2013(E), 2013
- BPSim 2013: Workflow Management Coalition: BPSim – Business Process Simulation Specification, Document Number WFMC -BPSWG-2012-1, 2013
- BPSim 2014: Workflow Management Coalition: BPSim Implementer’s Guide, 2014
- Cristina Cabanillas, Manuel Resinas, Antonio Ruiz-Cortés 2011: RAL: A High-Level User-Oriented Resource Assignment Language for Business Processes, BPM 2011 Workshops, LNBIP Vol 99, Springer, 2011
- Nick Russell, Wil M. P. van der Aalst, Arthur H. M. ter Hofstede, David Edmond 2005: Workflow Resource Patterns: Identification, Representation and Tool Support. CAiSE 2005: 216-232
- John Januszczak, Geoff Hook 2011: Simulation standard for business process management. Winter Simulation Conference 2011: 741-751
- Christian Müller et al. 2015A: Gegenüberstellung der Simulationsfunktionalitäten von Werkzeugen zur Geschäftsprozessmodellierung, TH Wildau, <http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:kobv:526-opus4-4354>
- Christian Müller, Klaus Bösing 2015b: Vergleich von Simulationsfunktionalitäten in Werkzeugen zur Modellierung von Geschäftsprozessen, in AKWI 2015, http://dx.doi.org/10.15771/978-3-944330-47-1_2015_1
- Raimar Scherer 2011: Process-Based Simulation Library for Construction Project Planning, Winter Simulation Conference 2011

Anthony Waller, Martin Clark, Les Enstone 2006: L-SIM: Simulating BPMN Diagrams with a Purpose Built Engine, Winter Simulation Conference 2006

AUTHORS BIOGRAPHIES



RALF LAUE studied mathematics at the University of Leipzig, Germany. After graduating, he worked as a system programmer before returning to the University of Leipzig in 2003. He obtained a PhD in computer science in 2010. Since 2011, he is a full professor for software engineering at the University of Applied Sciences in Zwickau, Germany. His research interests include the correctness and understandability of visual models in computer science.

His email address is: ralf.laue@fh-zwickau.de



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/> .

HYBRID MODEL OF HUMAN MOBILITY FOR DTN NETWORK SIMULATION

Alexander Privalov and Alexander Tsarev
Computer Science Department
Samara State Aerospace University
34, Moskovskoye shosse, Samara, 443086, Russia
E-mail: privalov1967@gmail.com

KEYWORDS

Human mobility models, wireless ad-hoc networks modeling.

ABSTRACT

A hybrid model of human mobility is presented. It combines features of the SLAW-type models and the Levy walk models to preserve advantages of the SLAW-type model in simulation of real human mobility and decrease computational time.

INTRODUCTION

An adequate model of node mobility in ad-hoc networks is very important for correct estimation of the network performance by simulation of the real networks' behaviour. Especially it is important for such class of ad-hoc networks as delay-tolerated networks (DTN). The DTN networks are characterized by small connectivity, so at some moment of time a connection between message sender and receiver may not exist and will appear only because of the nodes' positions changing. Therefore, adequateness of the mobility model to the real mobility is a key to the correct estimation of such fundamental characteristics of DTN networks protocols like probability of message delivery and probability distribution of transmission delay. It is the reason, why during the last decade a lot of efforts of research community were devoted to an investigation of real human mobility and to a development of adequate models of it.

Modern researches of human mobility reveal important features, like waypoint clustering and Levy-type distribution between consecutive waypoints (see, for example, (Brockmann et al. 2006; Gonzalez et al. 2008; Rhee et al. 2008; Rhee et al. 2011)). These features should be taken into account for correct modeling of user mobility in DTN networks. There are many modern well-known mobility models, like TLW (Rhee et al. 2011), CMM (Lim et al. 2006), ORBIT (Ghosh et al. 2007) and so on, which are able to catch some of the features of real human mobility, but not all of them.

Recently in (Lee et al. 2012; Lee et al. 2008) a new type of models was presented, which can be referred to as SLAW-type models (Self-Similar Least Action Walk) and are able simultaneously to catch several important features of real human mobility. Comparisons made in (Lee et al. 2012) show that these models outperform

models from (Rhee et al. 2011; Lim et al. 2006; Ghosh et al. 2007) in catching real mobility features. However, as we demonstrate in this report, SLAW-type models can take much computational time. This circumstance could be important, if these models are used for simulation of DTN networks with large number of nodes.

In this report we propose a hybrid model of human mobility, which combines features of the random Levy walks with some features of SLAW. This presented model keeps useful features of SLAW, but is more effective in terms of computational time.

SHORT DESCRIPTION OF SLAW-TYPE MODELS

SLAW-type models consider human mobility as transitions between so-called waypoints, where the human stops for a noticeable amount of time. It is well-known (see, for example (Rhee et al. 2011)), that the distance between consecutive waypoints has probability distribution close to the Levy distribution. Time stopped in waypoint has probability distribution close to the Levy distribution too. These features are captured well by the random Levy walk model (see, for example (Rhee et al. 2011)), but the clustering of waypoints is not captured by this model.

In the SLAW-type models, clusters can be taken from a real mobility traces. For this purpose, the real trace is processed for finding waypoints, and then for grouping these waypoints into clusters. By definition, the waypoint is the center of a circle with radius r of several meters (usually $r = 5\text{m}$) where a human spends inside more than T seconds (usually $T=30\text{sec}$). A cluster of waypoints is a rectangle that include transitive closure of waypoints which are not further from each other then the distance R (usually $R=100\text{m}$, also 50m and 250m are used). Also, SLAW-type models use self-similar parameter (variance) from the real trace (Lee et al. 2008).

When simulated trace is generated, its waypoints are randomly distributed inside clusters with the same self-similar parameter (variance), as in the real trace. As showed in (Lee et al. 2012), to provide Levy distribution of distances between consecutive waypoints, waypoints inside a cluster should be visited according LATP (Least Action Trip Principle), and after visiting all waypoints inside the current cluster, the

next cluster should be selected according LATP again (using distance between clusters).

According LATP, the next location to visit (as waypoint as cluster) is selected with probability, inversely to the distance to it in some power (parameter of the model). I.e. while the current location (waypoint or cluster) is i and the set of all locations is V , then the probability of selecting the next location with number j is calculated as follows:

$$\Pr\{i \rightarrow j\} = \frac{\left(\frac{1}{d_{ij}}\right)^p}{\sum_{k \in \{V-V'\}} \left(\frac{1}{d_{ik}}\right)^p} \quad (1)$$

where d_{ij} is the Euclidean distance from the location i to j , parameter p is a fixed real variable with values in the range $[0; +\infty)$, and V' is a set of locations from V , that have already been visited. Parameters p (there are two different parameters – one for waypoints selection, another one for clusters selection) can be used to fit the distribution of transition distances in simulated trace to the distribution in real trace.

In fact, for calculations according LATP, if there are N waypoints in cluster, it is necessary to calculate $N(N-1)/2$ distances between waypoints, i.e. $O(N^2)$ operations. This could require a large computational time, which the presented hybrid model can decrease.

HYBRID MODEL DESCRIPTION

As in SLAW-type models, proposed model uses clusters, which were found by the processing real traces data. At the beginning of the simulation run, each user gets a general information about its trace in the form of a set of clusters to be visited during the trip. User chooses first cluster and takes random points inside the cluster as a start position. Then it moves inside the cluster, making Levy steps and pauses after each step for Levy distributed time (the point of pause is the waypoint). Probability density of one Levy step is

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(-itx - |ct|^\alpha) dt \quad (2)$$

Usually the direction of simple Levy walk is uniformly distributed, but in our model, it is not the case. Instead of uniform distribution of the direction of next Levy step, the direction of the step is selected to prevent exit from the cluster as long as possible. If the next step in random direction is going to be outside of the cluster, this direction is changed directly toward the most distant corner of the cluster. Such change has maximal chance to keep the moving node (human) inside the cluster. If after the direction change the step is still outside the cluster, it means that it is time to change the

cluster and to select the next cluster. Transitions between clusters, like in SLAW-type models, go on according LATP. From the set of given at the beginning of simulation and still unvisited clusters, the user chooses the next cluster with the probability (1), and takes a random point inside it as end point for inter-cluster step. After transition to the new cluster user starts movement inside cluster as described above.

It is obvious, that calculation time of movement inside the cluster is proportional to the number of waypoints inside the cluster, i.e. $O(N)$. Therefore, for large N we can expect that this model will be faster than the SLAW-type model.

EXPERIMENTAL RESULTS

For comparison the results of real traces simulation by hybrid and SLAW-type models, both of them were implemented in the OMNET++ simulation system (Varga András 2001) with INET framework (Steinbach Till et al. 2011). Real traces of human mobility were taken from (Kotz D. 2015). These traces are the files with records of movement of one person (moving node). Each record is the time stamp with 30 sec step and two coordinates of the node at this time. Therefore, before using this data in our experiments, traces are processed for waypoints and clusters finding. A set of points from the file of real trace, which will join to the one waypoint, is determined as follow: at the beginning this set is empty, and then each point being successively read from the trace file is tested for possibility to be added to the set; if the new point and all members of the set are inside the circle of radius r , then the point is added to the set. If the point can't be placed in the circle with member of the set, then the current waypoint is complete, its coordinates are coordinates of the covering circle, and the new point is the first member of a set for new waypoint. After the transformation of the real trace into sequence of waypoints, the cluster bounds are detected according above cluster definition. Also the self-similar parameters (variances) necessary for SLAW-type model are calculated according formulas from (Lee et al. 2008).

To compare the simulated mobility with the real one the complementary cumulative distribution function (CCDF) of a distance between consecutive waypoints (including transitions between clusters) is used. By definition, CCDF is

$$\bar{F}(x) = \Pr\{X \geq x\} = 1 - F(x) \quad (3)$$

and after waypoint finding for the real trace, appropriate data are collected.

Then hybrid model with the same clusters as in the real trace run several times to find parameters c and α and parameter p of LATP cluster selection procedure, that provide closeness of CCDF for the trace simulated by hybrid model to CCDF of the real trace. In the

simulated trace, there are many small Levy steps, which are less than r , therefore described above procedure of

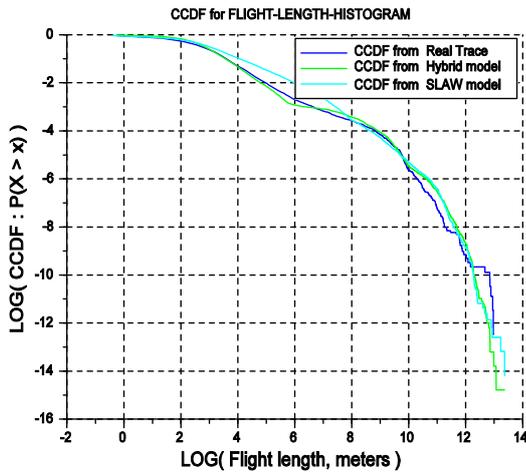


Figure 1: CCDF for KAIST

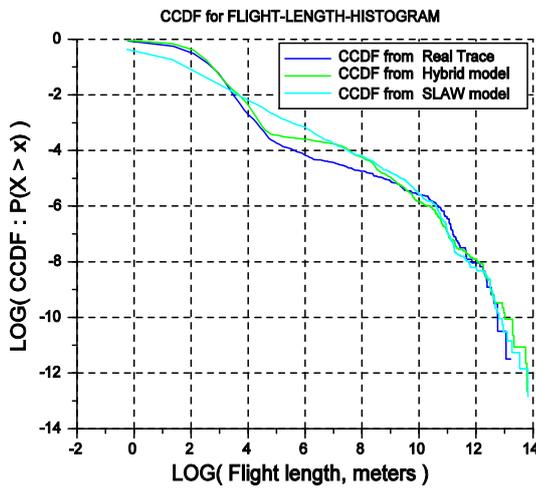


Figure 2: CCDF for NCSU

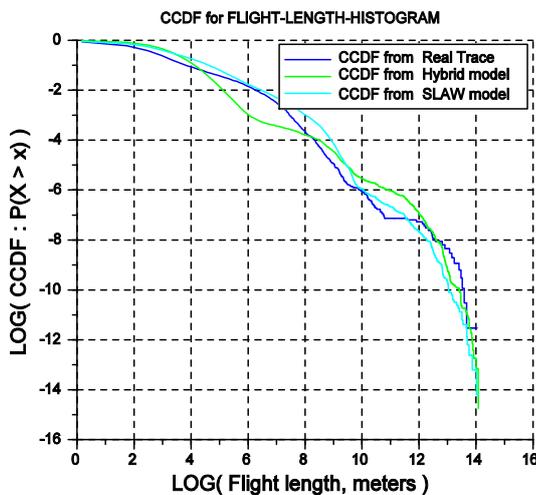


Figure 3: CCDF for Disney World

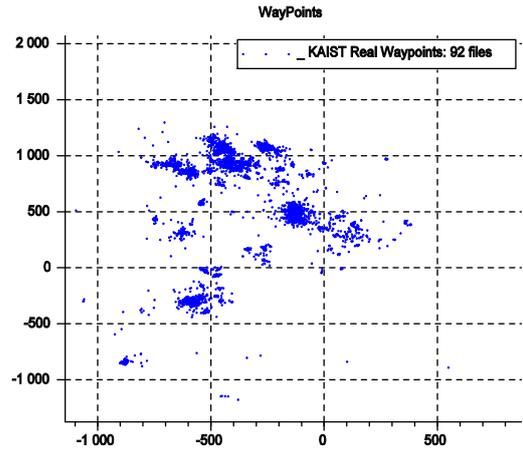


Figure 4: Waypoints for KAIST real trace

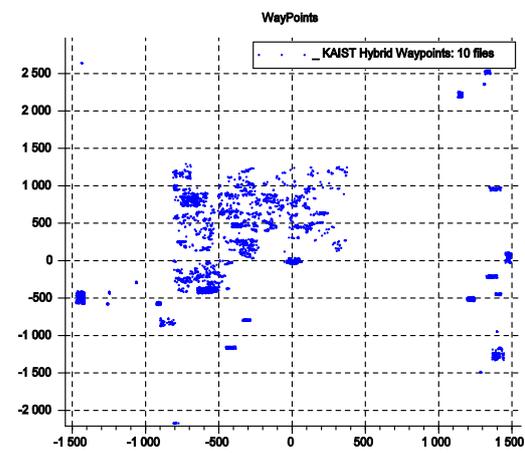


Figure 5: Waypoints for KAIST Hybrid model

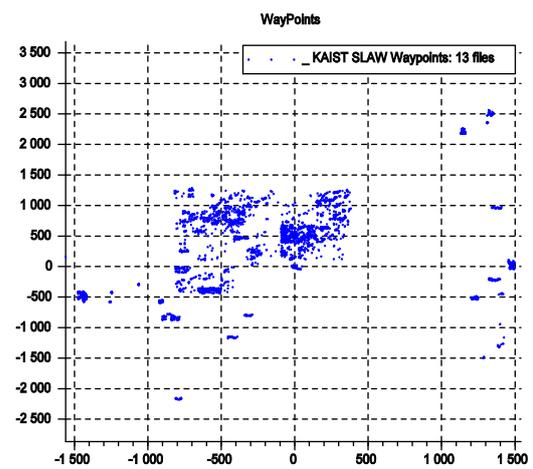
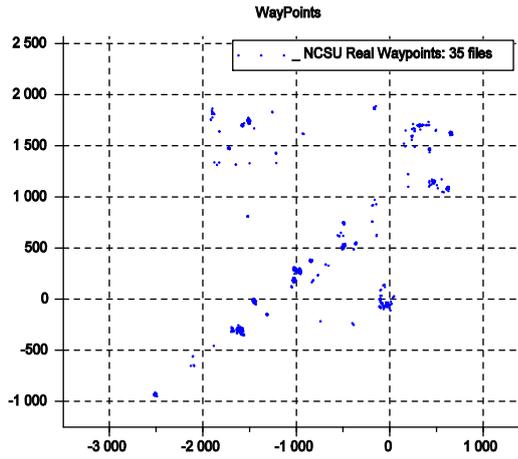
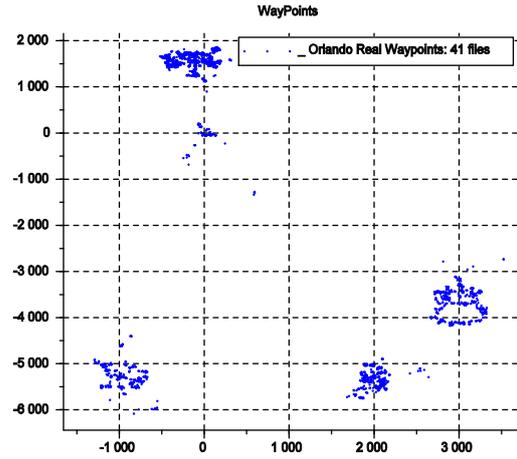


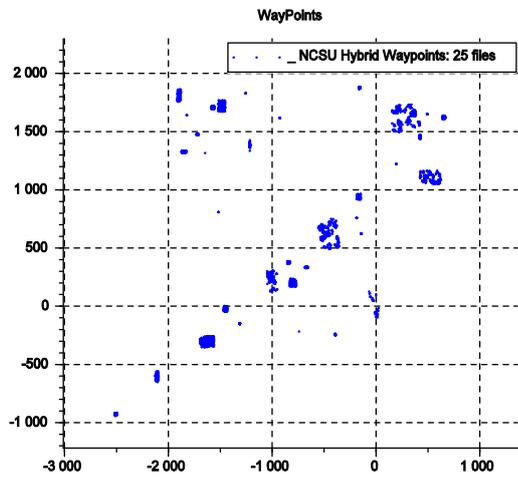
Figure 6: Waypoints for KAIST SLAW model



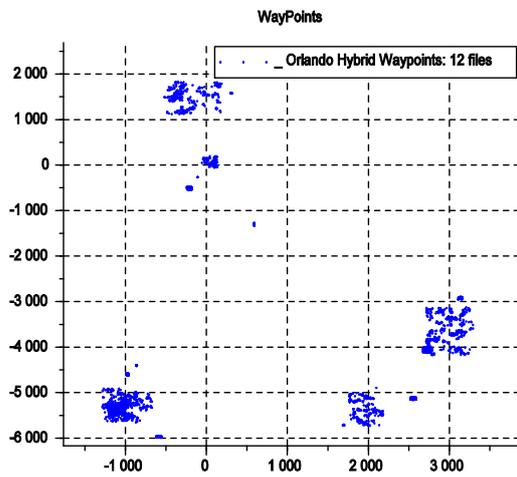
Figures 7: Waypoints for NCSU real trace



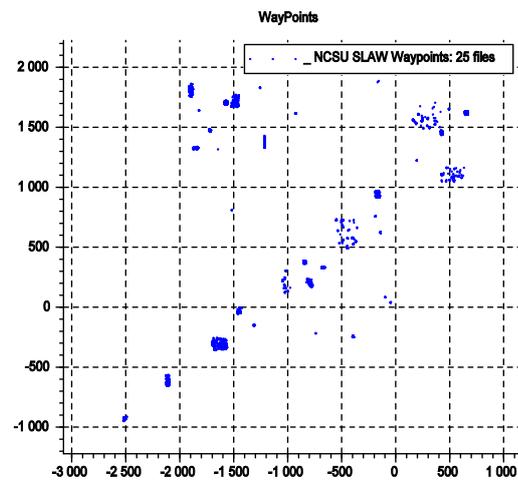
Figures 10: Waypoints for Disney real trace



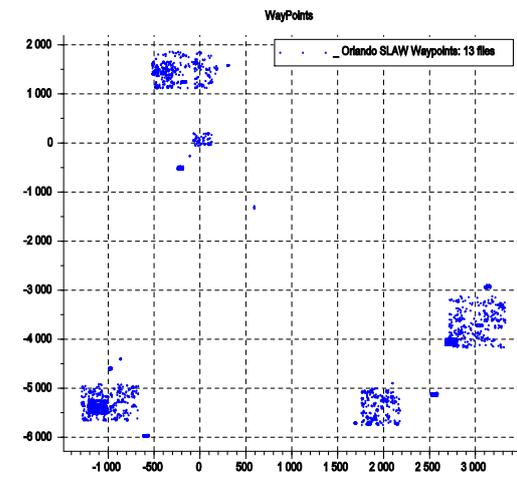
Figures 8: Waypoints for NCSU Hybrid model



Figures 11: Waypoints for Disney Hybrid model



Figures 9: Waypoints for NCSU SLAW model



Figures 12: Waypoints for Disney SLAW model

waypoints finding performs on simulated trace before the CCDF calculation.

The numbers of waypoints in each cluster of hybrid model trace are saved to use for SLAW-type model runs. According to description of SLAW model in (Lee et al. 2008), our version of SLAW model was developed. It intended just for comparison with hybrid model in computational complexity (computational speed). Our version of the SLAW model uses clusters and self-similar parameters from real trace and numbers of waypoints in each cluster from hybrid model. Given number of waypoints are distributed inside appropriate cluster in self-similar manner and then these points are visited in sequence according LAMP with some parameter p_1 . After visiting the last waypoint inside the cluster, next cluster is chosen according to LAMP with parameter p_2 . Just to complete the picture we run SLAW model several times to find appropriate p_1 and p_2 that provide closeness of the SLAW model CCDF to the CCDF of the real trace.

Our purpose in fitting CCDF of our version of SLAW model is just to demonstrate that this version is able to work. Real ability with deep details of SLAW-type models see in (Lee et al. 2012; Lee et al. 2008).

Here we present the results for three data sets from collection in (Kotz D. 2015): from KAIST (Korea Advanced Institute of Science and Technology), from NCSU (*North Carolina State University*) and from Disney World in Orlando. All these data sets were obtained by the same way: 50 volunteers (university students) during the day (or during the visit to the park) carried in the pocket GPS navigator, which recorded its position every 30 sec. We used these data to find real waypoints, real clusters of waypoints and other parameters for SLAW-type and Hybrid models.

Both models run for 50 mobile nodes, for model time of 5 days. The computational time is presented in Table I. On the figures 1-3, CCDF of the distance between consecutive waypoints are presented for all places. It is shown that CCDF of both models have about the same closeness to the CCDF of the real trace.

In addition, we present a set of pictures of waypoints positions for real trace and for both models to show clustering ability (figures 4-12). All pictures have approximately the same number of waypoints. For each dataset an only part of the whole area of simulation is presented to show clustering ability more clearly. It is clear, that waypoint distribution patterns for both models have about the same closeness to the pattern of the real trace.

Table 1: Computational time (sec)

Data Set	SLAW model	Hybrid model	Ratio
KAIST	391.4	167.4	2.34
NCSU	323.8	85.5	3.79
Disney	476.4	179.6	2.65

CONCLUSIONS

The hybrid model of human mobility is presented. This model captures such important features of real mobility as Levy steps between waypoints and waypoints clustering and is faster in simulation then the SLAW model. Therefore, the hybrid model can have advantages in simulation of DTN networks with large number of nodes.

REFERENCES

- Brockmann D; L. Hufnagel; and T. Geisel. 2006. "The scaling laws of human travel." *Nature*, vol.439 (Jan), pp.462-465.
- Gonzalez M.C.; C.A. Hidalgo; and A.-L.Barabasi. 2008. "Understanding individual human mobility patterns." *Nature*, vol.453 (Jun), pp.779-782.
- Rhee I.; M. Shin; S. Hong; K. Lee; and S. Chong. 2008. "On the Levy walk nature of human mobility." *Proc. IEEE INFOCOM* (Phoenix, AZ, Apr.) pp.924-932.
- Rhee I.; M. Shin; S. Hong; K. Lee; S.J. Kim; and S. Chong. 2011. "On the Levy-walk nature of human mobility." *IEEE/ACM Trans. on Networking*, vol.19, №3 (June), pp.630-643.
- Lim S.; C.Yu; and C.R.Das. 2006. "Clustered mobility model for scale-free wireless networks." *Proc. IEEE LCN 2006* (Tampa, FL, Nov.) pp. 231 – 238.
- Ghosh J.; S.J. Philip; and C. Qiao. 2007. "Sociological orbit aware location approximation and routing (solar) in MANET." *Ad hoc Netw.* Vol.5, pp. 189-209.
- Lee K.; S. Hong; S.J. Kim; I. Rhee; and S. Chong. 2012. "SLAW: Self-Similar Least-Action Human Walk." *IEEE/ACM Trans. on Networking*, vol.20, №2 (Apr), pp.515-529.
- Lee K.; S. Hong; S.J. Kim; I. Rhee; and S. Chong. 2008. "Demystifying Levy Walk Patterns in Human Walks." *Technical Report* (CSC, NCSU) URL: http://research.csc.ncsu.edu/netsrv/sites/default/files/Demystifying_Levy_Walk_Patterns.pdf
- Varga András. 2001. "The OMNeT++ discrete event simulation system." *Proceedings of the European simulation multiconference (ESM'2001)*. Vol. 9. No. S.185.sn. (OMNET++ community cite: <http://omnetpp.org> (access date 19.03.2016))
- Steinbach Till; et al. 2011. "An extension of the OMNeT++ INET framework for simulating real-time ethernet with high accuracy." *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques. ICST*.
- Kotz D. 2015. "Community Resource for Archiving Wireless Data at Dartmouth." *Dartmouth College*. (URL: <http://www.crawdad.org/index.html> (access date 19.03.2016))

AUTHOR BIOGRAPHIES

ALEXANDER YU. PRIVALOV received the M.S. degree from Moscow Institute of Physics and Technology, Moscow, Russia, in 1990, and the Ph.D. degree from the Institute of Information Transmission Problems, Moscow, Russia, in 1993. He was with the Department of Technical Cybernetics, Samara State Aerospace University, as an Associate Professor. In 1996-1998, he has been a Visiting Research Scholar in computer science and telecommunications with the University of Missouri-Kansas City. Since 1998, he was with Samara State Aerospace University again, since 2004 as Full Professor and since 2013 as Chief of the Chair of Applied Mathematics. His research interests include modeling and performance evaluation of

communication networks. His e-mail address is privalov1967@gmail.com.

ALEXANDER A. TSAREV (b. 1991) received master's degree of Information Technology in Samara State Aerospace University in 2014. Master's program is "Technology of parallel programming and supercomputing". Now he is postgraduate student in the department of applied mathematics in Samara State Aerospace University. E-mail: al-xandr1@yandex.ru.

OPTIMIZATION OF A HEAT RADIATION INTENSITY AND TEMPERATURE FIELD ON THE MOULD SURFACE

Jaroslav Mlynek
Roman Knobloch
Department of Mathematics , FP
Technical University of Liberec
Studentská 2, 461 17 Liberec,
Czech Republic
E-mail: jaroslav.mlynek@tul.cz
E-mail: roman.knobloch@tul.cz

Radek Srb
Institute of Mechatronics and
Computer Engineering
Technical University of Liberec
Studentská 2, 461 17 Liberec,
Czech Republic
E-mail: radek.srb@tul.cz

KEYWORDS

Intensity of heat radiation, optimization of temperature field, differential evolution algorithm, parallel programming, software implementation.

ABSTRACT

This article is focused on the infrared heating of shell metal moulds and optimization of temperature field on the surface of the mould. The upper part of the mould is heated by infrared heaters, and after the required temperature is attained the inner part of the mould is sprinkled with special PVC powder. The moulds are made of aluminium or nickel alloys. The described mathematical model allows us to optimize locations of heaters above the mould and thus we get an approximately uniform temperature field on the whole inner mould surface. In this way the whole surface of produced artificial leather has the same material structure and colour shade. A differential evolution algorithm is used to optimize the locations of heaters. A practical example of the optimization of the heaters locations and the calculation of the temperature field on the inner part of mould surface is included at the end of the article. The described process is one of the economical ways of artificial leathers production in the car industry.

INTRODUCTION

This article concerns economically beneficial technology for producing artificial leathers in the automotive industry. The artificial leathers are used as final parts for some components of car interior equipment (e.g. the inner surface of the door padding, the surface of the dashboard).

In practice, a metal mould is at first preheated by infrared heaters located above the mould. Then the inner mould surface is sprinkled with a special PVC powder and the upper part of mould surface is continually heated for about 3 minutes to an approximate temperature of 250°C.

The infra heaters have a tubular shape and their length is between 15 and 30 cm (see Figure 1). The heater is equipped with a mirror located above the radiating tube, which reflects heat radiation in the adjusted direction. Shell metal moulds of different proportions with

variously complicated surfaces and with weights approximately up to 300 kg are used in the production of artificial leathers (see Figure 2). The shell mould thickness is constant. For various moulds this constant thickness is from 6 to 8 mm. Moulds are usually made of aluminium or nickel alloys.



Figure 1: Ushio infra heater with 1000 W power

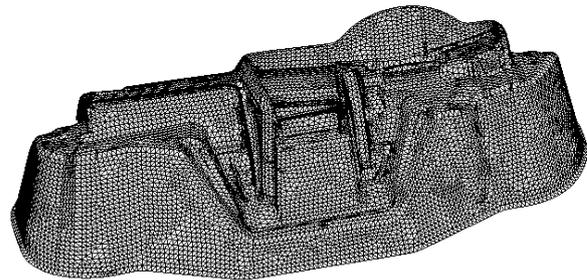


Figure 2: Aluminium mould of a passenger car dashboard

An important requirement for manufacturing artificial leathers is approximately uniform temperature of the whole inner mould surface in the given time during heating of the mould. In this way we obtain the required uniform material structure and colour shade of the artificial leather surface.

The above mentioned requirement of the uniform temperature can be fulfilled by finding suitable locations of heaters over the mould. The aim is to optimize the locations of the heaters, that is to find such locations that ensure the uniform radiation on the whole heated surface of the mould. Up to now, suitable locations of heaters were determined by manufacturing technicians based on their operational experience. Nevertheless, this approach did not provide a sufficiently uniform temperature and was time consuming.

The infrared heaters above the mould are usually located at a distance of between 5 and 30cm over the mould surface. In our model we also do not know the heater

distribution intensity function from the heater manufacturer. Therefore, experimental values for the heat radiation intensity by a sensor in the surroundings of the infra heater were measured. For radiation intensity calculations it was necessary to use transformations of Cartesian coordinate systems in Euclidean space E_3 and the linear interpolation at individual points of the mould surface.

During the optimization process of heaters locations we have to take into account the possible collisions of two heaters as well as collisions of a heater and the mould surface. Therefore, the optimization process is more complicated. The minimized function has many local extremes and for this reason it is not suitable to use gradient methods in the optimization. We used an evolutionary algorithm, specifically the differential evolution algorithm (using this algorithm, we obtain better results than when using a genetic algorithm, see (Mlynek and Srb 2012)). Evolutionary algorithms in general require a lot of operations and long computation time (especially when the mould volume is larger and we use a higher number of heaters). This is the main reason for us to use parallel programming techniques in the optimization process. The whole optimization process was programmed by the authors in the Matlab system and we used the Matlab Parallel Computing Toolbox.

For finding optimized locations of the heaters and the optimized heat radiation intensity on the whole irradiated surface of the mould we calculate temperature evolution (during the heating of the mould surface) on the inner part of the mould surface. We solve the partial differential parabolic equation of the heat conduction in the mould with initial and boundary conditions. We use the ANSYS software package to obtain the solution to this equation.

The used mathematical model, the process of calculation of the heat radiation intensity on the mould surface are described in more details in (Mlynek and Srb 2012) and the optimization process in (Mlynek and Srb 2014).

MATHEMATICAL MODEL OF HEAT RADIATION ON THE MOULD

We describe a heat radiation model in this chapter. The heaters and mould are represented in a 3-dimensional Euclidean space E_3 using a Cartesian coordinate system (O, x_1, x_2, x_3) , with basis vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$.

Representation of a heater

The heater is represented by a straight line segment with a given length (see Figure 3). The position of every heater H can be defined by the following 6 parameters

$$H : (s_1, s_2, s_3, u_1, u_2, \varphi), \quad (1)$$

where the first three parameters are the coordinates of the heater centre S , the following two parameters are the

first two coordinates of the unit vector \mathbf{u} of the heat radiation direction (the third coordinate is negative, i.e. the heater radiates “downward”). The last parameter is the angle φ between the vertical projection of vector \mathbf{r} of the heater axis onto the x_1x_2 -plane and the positive part of axis x_1 ($0 \leq \varphi < \pi$, the vectors \mathbf{u} and \mathbf{r} are orthogonal).

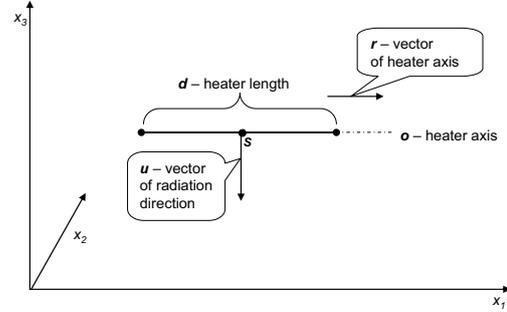


Figure 3: Schematic representation of the heater

Representation of a mould

The upper mould surface P_{up} is described by elementary surfaces p_j , where $1 \leq j \leq N$. It holds that $P_{up} = \cup p_j$, where $1 \leq j \leq N$ and $\text{int } p_i \cap \text{int } p_j = \emptyset$ for $i \neq j$, $1 \leq i, j \leq N$. Each elementary surface p_j is described by its centre of gravity $C_j = [c_1^j, c_2^j, c_3^j]$, by the unit outer normal vector $\mathbf{v}_j = (v_1^j, v_2^j, v_3^j)$ at the point C_j (we suppose \mathbf{v}_j faces “upwards” and therefore is defined through the first two components v_1^j, v_2^j) and by the area of elementary surface w_j . Therefore each elementary surface p_j can be defined by the following 6 parameters

$$p_j : (c_1^j, c_2^j, c_3^j, v_1^j, v_2^j, w_j). \quad (2)$$

Calculation of total heat radiation

Now, we describe the numerical procedure for computation of total heat radiation intensity on the upper mould surface P_{up} . We denote L_j as the set of all heaters radiating on the j -th elementary surface p_j ($1 \leq j \leq N$) for the fixed position of heaters, and I_{jl} the heat radiation intensity of the l -th heater on the p_j elementary surface (in our model we suppose I_{jl} is constant on the whole p_j and is calculated in its centre of gravity C_j). Then the total radiation intensity I_j on the elementary surface p_j is given by the following relation (more details in (Cengel 2007))

$$I_j = \sum_{l \in L_j} I_{jl}. \quad (3)$$

A detailed description of the procedure for calculating

the heat radiation intensity I_{jl} of the l -th heater in a general position at the elementary surface p_j is shown in (Mlynek and Srb 2012).

OPTIMIZATION OF HEAT RADIATION INTENSITY

In this chapter we describe a procedure for optimizing the location of heaters. We need to determine such location of heaters that heat radiation intensity on the whole upper mould surface P_{up} will be approximately equal to the constant value of heat radiation intensity I_{rec} recommended by the producer of artificial leathers.

We can define the deviation function F (respectively \tilde{F}) of heat radiation intensity by the relations

$$F = \frac{1}{W} \cdot \sum_{j=1}^N |I_j - I_{rec}| w_j, \quad (4)$$

$$\tilde{F} = \left(\frac{1}{W} \cdot \sum_{j=1}^N (I_j - I_{rec})^2 w_j \right)^{1/2}, \quad (5)$$

where $W = \sum_{j=1}^N w_j$. We highlight that w_j denotes the area of the elementary surface p_j . Our goal is to find the minimum of function F (respectively \tilde{F}). Function F (and analogously function \tilde{F}) has many local minima. Using the gradient method for minimizing function F (respectively \tilde{F}) is not appropriate because there is a high probability of failure of this method (finding only a local inconvenient minimum). Therefore, we use an evolutionary algorithm to find a minimum of function F (respectively \tilde{F}). In accordance with relation (1), the location of every heater H is defined by 6 parameters. Let us suppose we use M heaters for heating of the mould. Therefore, $6M$ parameters are necessary to define the positions of all M heaters. We successively construct populations of individuals y by the evolutionary algorithm (for more details see (Price et al. 2005)). Every individual y represents one possible location of heaters above the mould.

We seek the individual $y_{\min} \in C$ satisfying the condition

$$F(y_{\min}) = \min\{F(y); y \in C\}, \quad (6)$$

where $C \subset E_{6M}$ is the examined set. Every element of C is formed by a set of $6M$ allowable parameters and this set defines just one location of heaters above the mould. We keep track of the following three types of collisions during generation of individuals y : a heater radiates more on another heater than the given limit, a heater has an insufficient distance from another heater, a heater has an insufficient distance from the mould surface. If a collision occurs then the corresponding individual y is penalized and excluded from the population.

The finding of the individual y_{\min} defined by relation (6) is not realistic in practice. However, we are able to

determine an optimized solution y_{opt} that is satisfactory for the production requirements.

Every population of our evolutionary algorithm includes NP individuals. The generated individuals are saved in the matrix $\mathbf{B}_{NP \times (6M+1)}$. Every row of this matrix represents one individual y , and its evaluation $F(y)$.

Now, we briefly describe the used differential evolution algorithm.

Differential evolution algorithm

Our minimization problem is solved by the differential evolution algorithm named *DE/rand/1/bin* (for more details see (Price et al. 2005)) and was programmed in the Matlab code by the authors. We describe schematically the individual steps of the algorithm.

Steps of the differential evolution algorithm

Input: The initial individual y_1 , population size NP , the number of used heaters M (dimension of the problem is $6M$), crossover probability CR , mutation factor f , the number of calculated generations NG .

Internal computation:

1. create an initial generation ($G=0$) of NP individuals y_i^G , $1 \leq i \leq NP$,

2. a) evaluate all the individuals y_i^G of the generation G (calculate $F(y_i^G)$ for every individual y_i^G), b) store the individuals y_i^G and their evaluations $F(y_i^G)$ into the matrix \mathbf{B} ,

3. repeat until $G \leq NG$

a) for $i:=1$ step 1 to NP do

(i) randomly select index $k_i \in \{1, 2, \dots, 6M\}$,

(ii) randomly select indexes $r_1, r_2, r_3 \in \{1, \dots, NP\}$, where $r_l \neq i$ for $1 \leq l \leq 3$; $r_1 \neq r_2$, $r_1 \neq r_3$, $r_2 \neq r_3$;

(iii) for $j:=1$ step 1 to $6M$ do

if $(\text{rand}(0,1) \leq CR \text{ or } j=k_i)$ then

$$y_{i,j}^{trial} := y_{r_3,j}^G + f(y_{r_1,j}^G - y_{r_2,j}^G)$$

else

$$y_{i,j}^{trial} := y_{i,j}^G$$

end for (j)

(iv) if $F(y_{i,j}^{trial}) \leq F(y_{i,j}^G)$ then $y_i^{G+1} := y_i^{trial}$

else $y_i^{G+1} := y_i^G$

end for (i)

b) store individuals y_i^{G+1} and their evaluations $F(y_i^{G+1})$ ($1 \leq i \leq NP$) of the new generation $G+1$ into the matrix

\mathbf{B} ; $G := G+1$

end repeat.

Output: The row of matrix \mathbf{B} that contains the corresponding value $\min\{F(y_i^G); y_i^G \in \mathbf{B}\}$ represents the best found individual y_{opt} .

Comment: function $\text{rand}(0,1)$ randomly chooses a number from the interval $\langle 0,1 \rangle$. The notation $y_{i,j}^G$

means the j -th component of an individual y_i^G in the G -th generation. The individual y_{opt} is the final solution and includes information about the location of each heater H in the form (1).

CALCULATION OF TEMPERATURE FIELD IN THE MOULD

By using the differential evolution algorithm described in chapter "Optimization of Heat Radiation Intensity" we obtain the optimized individual y_{opt} . This means we receive suitable locations of heaters above the mould (each heater is defined in the form of relation (1)) and approximately uniform heat radiation intensity (close to the value I_{rec}) on the upper surface P_{up} of the mould (we know heat radiation intensity I_j for each elementary surface $p_j \in P_{up}$).

Now we describe the calculation of the temperature field in the mould for time of mould heating and for locations of heaters defined by optimized individual y_{opt} .

Mathematical model of the heat conduction

In this section we solve the parabolic equation of the heat conduction

$$c \rho \frac{\partial T(x,t)}{\partial t} = \lambda \Delta T(x,t) + Q(x,t) \quad (7)$$

on the domain $\Omega \subset E_3$, where Ω represents the shell mould. The function $T(x,t)$ denotes a temperature field in relation (7), point $x = (x_1, x_2, x_3) \in \Omega$, time $t \in (0, \tau)$, where τ is the duration of the heat radiation. The symbol Δ stands for the Laplace operator with respect to space variables, i.e.

$$\Delta T(x,t) = \sum_{i=1}^3 \frac{\partial^2 T(x,t)}{\partial x_i^2}.$$

The values c and ρ stand for the specific heat and mass density of the mould material. The value λ denotes the heat conductivity of the mould material. We suppose a homogeneous and isotropic material of the mould. The function $Q(x,t)$ represents the volume density of the heat sources. Nevertheless, in our case there is no inner heat source in Ω and so $Q(x,t) = 0$ in relation (7).

We consider the initial condition

$$T(x,0) = T_0 \quad \forall x \in \Omega, \quad (8)$$

where T_0 denotes the initial temperature of the mould. We choose the Newton boundary condition which suits best the situation when the hot body is surrounded by an environment (air) and when the heat transfer between the body and environment is possible. The simple Newton boundary condition can be expressed in the form

$$\lambda \frac{\partial T(x,t)}{\partial \nu} = -\alpha (T(x,t) - T_{air}),$$

where α is the coefficient of the heat transfer between the mould material and air, T_{air} is air temperature and ν is the unit vector of the outer normal. Nevertheless, this linear formulation of the boundary condition is not sufficient for our problem. There are two main reasons: 1/ the temperature of the mould reaches up to 300°C and it is not possible to neglect the own heat radiation of the mould determined by Stefan-Boltzmann law; 2/ there are no volume heat sources $Q(x,t)$ in the body of the mould and the heat is supplied exclusively by infrared heaters through the upper part P_{up} of the mould surface $\partial\Omega$. When we consider these two facts, we obtain the following form of the boundary condition (see e.g. (Incropera 2007))

$$\lambda \frac{\partial T(x,t)}{\partial \nu} = -\alpha (T(x,t) - T_{air}) - \varepsilon \sigma (T^4(x,t) - T_{air}^4) + I(x) \quad (9)$$

on the upper part P_{up} of the mould surface $\partial\Omega$ and for all other parts of the mould surface $\partial\Omega - P_{up}$

$$\lambda \frac{\partial T(x,t)}{\partial \nu} = -\alpha (T(x,t) - T_{air}) - \varepsilon \sigma (T^4(x,t) - T_{air}^4). \quad (10)$$

Here value ε denotes the emissivity of the mould and σ denotes the Stefan-Boltzmann constant, $\sigma = 5,775 \cdot 10^{-8} \text{ Wm}^2\text{K}^{-4}$. Value $I(x)$ indicates the heat radiation intensity on the part of mould surface P_{up} . In our case the value $I(x)$ is defined by locations of the heaters given by optimized solution y_{opt} from chapter "Optimization of Heat Radiation Intensity". The equation (7) together with initial condition (8) and boundary conditions (9) - (10) describes the heat conduction in the mould Ω . It is a nonstationary heat conduction problem, time τ of duration of heat radiation is usually 3 minutes. For needs of the production it is most important to know the temperature on the inner part of the mould surface P_{in} (the artificial leather is produced on this part of the mould surface).

Use of software package ANSYS

We use the ANSYS 15.0.7. software package to determine the solution of the parabolic equation of the heat conduction (7) with conditions (8) - (10). The input parameters for the ANSYS system are the mould surface described by triangle elementary surfaces in Euclidean space E_3 , heat radiation intensity on each elementary surface and thickness of the shell mould. The finite element method is used to solve our thermal problem by the ANSYS system. The corresponding base functions are quadratic.

PRACTICAL EXAMPLE

We find the optimized locations of the infrared heaters over the mould, and we calculate the temperature of the mould (especially the temperature on the inner surface P_{in} of the mould).

I. Input parameters

a/ heaters

type of infrared heater: Ushio, length 15 cm, width 4 cm, power 1000 W, all the heaters are of the same power and shape, number of heaters: 16,

b/ mould

mould size: $0,9 \times 0,4 \times 0,15 \text{ m}^3$, mould thickness: 8 mm, material of the mould: aluminium alloy, number of the elementary surfaces of the mould $N = 2178$, specific heat $c = 875 \text{ J/kgK}$, mass density $\rho = 2770 \text{ kg/m}^3$, heat conductivity $\lambda = 160 \text{ W/mK}$, coefficient of the heat transfer between the mould material and air $\alpha = 20 \text{ W/m}^2\text{K}$, emissivity of the mould $\varepsilon = 1$,

c/ heat radiation

recommended heat radiation intensity I_{rec} by producer of the artificial leather: $I_{rec} = 47 \text{ kW/m}^2$, duration of the heat radiation $\tau = 180 \text{ s}$,

d/ temperature

temperature of air $T_{air} = \text{initial temperature } T_0 = 22 \text{ }^\circ\text{C}$.

II. Starting locations of the heaters in optimization process

The heaters in the initial locations lie in the plane given by axes x_1 and x_2 and at a distance of 10 cm over the mould top and with heater axis r parallel to axis x_1 . Initial locations of heaters over the mould are represented by individual y_1 and corresponding heat radiation intensity on the upper part P_{up} of mould surface is displayed in Figure 4. The value of the function F (defined by relation (4)) for the initial individual y_1 is $F(y_1) = 20,87$.

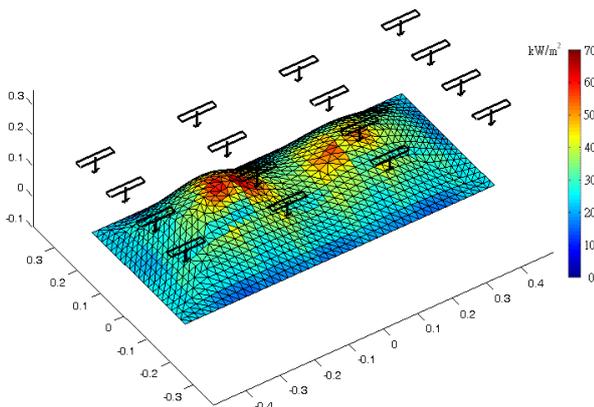


Figure 4: The initial locations y_1 of the heaters and heat radiation intensity I (kW/m^2) on the mould surface.

III. Results

a/ optimized locations of heaters and heat radiation intensity

We receive the optimized individual y_{opt} by the differential evolution algorithm described in section “Steps of the differential evolution algorithm” with the following input parameters: crossover factor $CR = 0,98$; mutation factor $f = 0,60$; population size $NP = 200$; number of calculated generations $NG = 10\,000$. The locations of heaters over the mould corresponding to individual y_{opt} and appropriate heat radiation intensity are shown in Figure 5. Deviation function F (defined by relation (4)) has for individual y_{opt} value $F(y_{opt}) = 2,30$.

b/ optimized locations of heaters and heat conduction

We solve the parabolic equation of the heat conduction (7) with conditions (8) – (10) for heat radiation intensity on the upper part of the mould surface P_{up} corresponding to initial individual y_1 , y_{opt} and for the case when the heat radiation intensity has the constant value $I = I_{rec}$ (recommended heat radiation intensity by the producer of artificial leathers).

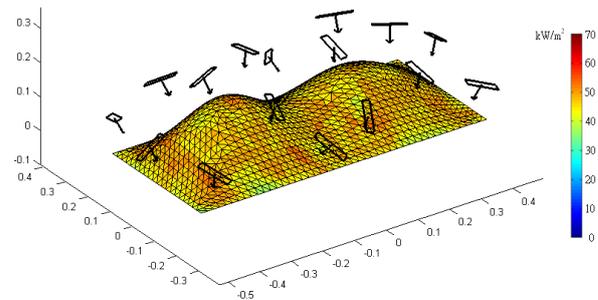


Figure 5: The locations of heaters and heat radiation intensity I corresponding to the individual y_{opt} .

For producers of leathers the most important knowledge is the temperature field on the inner part of the mould surface P_{in} (here the artificial leather is gradually formed and it is necessary that the deviations of temperatures at specific time of warming are less than 15°C). The temperature field at the heating time $t = 180 \text{ s}$ on the inner part P_{in} of the mould surface for above three mentioned cases are successively displayed in Figures 6, 7 and 8.

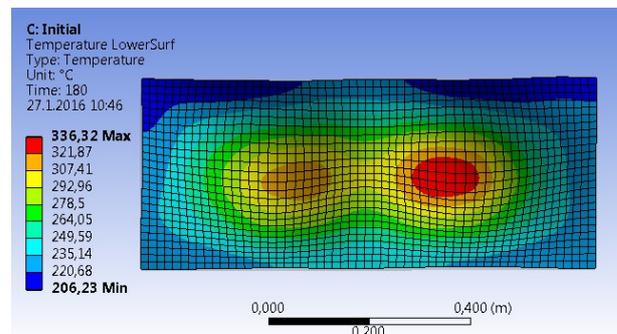


Figure 6: Temperature field corresponding to initial individual y_1 .

Furthermore, we determine the temperature deviation D on the inner part of the mould surface P_{in} . The value of D is defined by the following relation

$$D(y, t_D) = \sqrt{\int_{P_{in}} (T(x, t_D, I_y) - \bar{T}(t_D, I_y))^2 dx} \cong \quad (11)$$

$$\cong \sqrt{\sum_{j=1}^N (T(C_j, t_D, I_y) - \bar{T}(t_D, I_y))^2 w_j} .$$

Here $T(x, t_D, I_y)$ denotes the temperature at point $x \in P_{in}$ at heating time t_D for heat radiation intensity I_y corresponding to individual y . The term $\bar{T}(t_D, I_y)$ represents the average temperature on the surface part P_{in} for the given time t_D and heat radiation intensity I_y . We remind that C_j denotes the centre of gravity of elementary surface p_j and w_j is the area of this elementary surface.

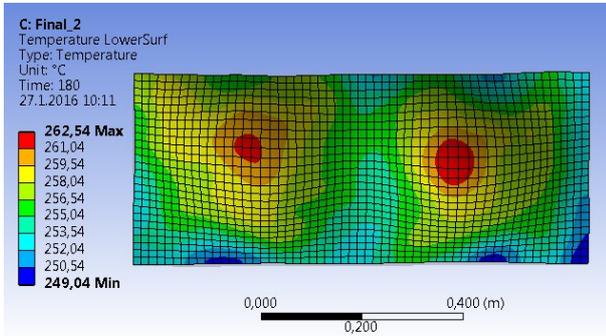


Figure 7: Temperature field corresponding to optimized individual y_{opt} .

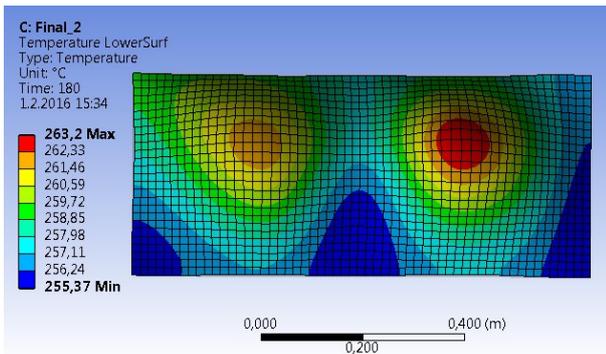


Figure 8: Temperature field corresponding to the recommended constant heat radiation intensity I_{rec} on the upper part of the mould surface.

We calculated deviations $D(y_1, t_D)$, $D(y_{opt}, t_D)$ and $D(y_{rec}, t_D)$, where y_{rec} denotes a virtual individual corresponding to the recommended fully homogeneous heat radiation intensity I_{rec} on the mould surface part P_{up} . We computed these values for $t_D = 180s$:

$$D(y_1, 180) = 19,332$$

$$D(y_{opt}, 180) = 1,522$$

$$D(y_{rec}, 180) = 1,135.$$

The last two values demonstrate that the optimized solution y_{opt} differs only very little from the ideal solution y_{rec} and is therefore fully acceptable for the practical production needs.

CONCLUSIONS

The mentioned technology of artificial leather manufacturing in the automotive industry is economically feasible (low energy consumption during production). The suitable locations of the heaters above the mould and their connection to an auxiliary constructions has been up to now conducted by experienced technicians. However, this approach was very laborious and time consuming (approximately 2 weeks). Furthermore, the optimized solution obtained by using our mathematical model is much more accurate and the calculation time when using parallel programming tool is on average approximately 4 hours.

We are now able to determine the temperature of the mould during its heating for the optimized locations of the heaters. This information is important for the manufacturer of artificial leathers, but it was not available before. Only temperature sensors were up to now positioned at some points of the mould. Knowledge of the temperature is important in particular on the inside surface part of the mould (where the artificial leather is produced). The temperature differences on this part of the mould surface have to be maintained at less than 15°C at the given moment of the mould heating. Such a temperature distribution ensures a uniform material structure and colour shade on the whole leather surface.

We obtained better results using the differential evolution algorithm than using a genetic algorithm (Mlynek and Srb 2012) in numerical experiments.

The described mathematical model and algorithm allows us to find the optimized locations of the heaters over the mould and in this way to determine the corresponding heat radiation intensity on the irradiated mould surface and to calculate the temperature field on the inner surface of the mould. In addition, using our mathematical model and calculation does not add any additional economic costs for the manufacturer.

ACKNOWLEDGEMENTS

This work was supported by grants SGS-FP-TUL and SGS-FM-TUL.

REFERENCES

- Cengel, Y. 2007. "Heat and Mass Transfer". McGraw-Hill, New York, 61-130, ISBN 9780070634534. 663-772.
- Cook, S. 2013. "CUDA programming: a developer's guide to parallel computing with GPUs". Elsevier, Walrhan, USA, ISBN 978-0-12-415933-4. 305-404.
- Incropera, F. P., D. P. De Witt, T. L. Bergman and A. S. Lavine. 2007. "Fundamentals of Heat and Mass Transfer".

John Wiley & Sons, Inc., ISBN-13: 978-0471457282. 8 – 12.

Mlynek, J. and R. Srb. 2012. "The Process of Optimized Heat Radiation Intensity Calculation on a Mould Surface". In *Proc. of the 26th European Conference on Modelling and Simulation*, K. G. Troitzsch, M. Möhring and U. Lotzmann (Eds.), May 2012, Koblenz, Germany, ISBN 978-0-9564944-4-3, DOI: 10.7148/2012-0461-0467. 461-467.

Mlynek, J. and R. Srb. 2014. "Differential Evolution and Heat Radiation Intensity Optimization". In *Proc. of Conf. Mathematics and Computers in Sciences and Industry*, M. Mastorakis (Ed.), Varna, Bulgaria, IEE Computer Society, ISBN 978-1-4799-4744-7, DOI 10.1109/MCSI.2014.11.135-138.

Price, K. V., R. M. Storn and J. A. Lampien. 2005. "Differential Evolution". Springer-Verlag, Heidelberg, ISBN 978-3-540-20950-8. 37-187.

AUTHOR BIOGRAPHIES

JAROSLAV MLYNEK was born in Trnava, Czechoslovakia and went to the Charles University in Prague, where he studied numerical mathematics at the Faculty of Mathematics and Physics and he graduated in 1981. In his work he focuses on the computational problems of heating and thermal losses in components of electrical machines and on mathematical models of thermal convection in electric machines. Currently he

works as an associate professor at the Technical University of Liberec, the Czech Republic. His e-mail address is: jaroslav.mlynek@tul.cz.

ROMAN KNOBLOCH was born in Turnov, the Czech Republic. He completed his studies at the Charles University in Prague, the Faculty of Mathematics and Physics where he studied scientific physics and teaching of mathematics and physics. His main areas of interest are: modelling of physical phenomena, stationary and non stationary heat processes and heat and transport phenomena in continuum mechanics. He works as an assistant professor at the Technical University of Liberec where he also participates in the PhD study programme. His e-mail address is roman.knobloch@tul.cz.

RADEK SRB was born in Mladá Boleslav, the Czech Republic and went to the Technical University of Liberec, the Czech Republic, where he studied computer science and programming at the Faculty of Mechatronics. He graduated in 2005. He focuses on problems concerning the automated control of production. He works as a teacher and he is a PhD student. His e-mail address is: radek.srb@tul.cz.

SOCIAL AND ECOLOGICAL CAPABILITIES FOR A SUSTAINABLE HIERARCHICAL PRODUCTION PLANNING

Marco Trost
 Prof. Dr. Thorsten Claus
 Enrico Teich
 Maximilian Selmair
 Department of Business Science
 Dresden Technical University
 Markt 23, 02763 Zittau, Germany
 E-mail: marco.trost@mailbox.tu-dresden.de

Prof. Dr. Frank Herrmann
 Innovation and Competence Centre for Production
 Logistics and Factory Planning (IPF)
 Regensburg Technical University of Applied Sciences

KEYWORDS

production planning, hierarchical planning, social variables, ecological variables, sustainable production planning, sustainable hierarchical production planning

ABSTRACT

Production planning and production control mainly focus on optimising the entire production system of a company. On the basis of hierarchical planning as a suitable method for solving this task, this paper shows - besides the economic dimension taken into account so far - that there are also social and ecological effects which will have to be considered in the process of planning. For this purpose, we would like to indicate here which social and ecological parameters can be or have already been taken into account for master production scheduling, for lot sizing and resource scheduling. As a result, an overview has been created which presents the existing concepts of sustainable production planning and production control as well as the existing deficits regarding the sustainability perspective.

INTRODUCTION

The concepts of hierarchical planning, as put forward by Hax, and Meal (1975), represent the state of the art in research and industry. In this paper, we have looked at hierarchical planning considering the restricted capacities as suggested by Drexel et al. (1994), with respect to production planning and production control (PPC). We distinguish between three planning stages: Master production scheduling (1), lot sizing (2) and resource scheduling (3) (refer for figure 1).

The planning approaches applied so far mostly consider single stages of planning and forbear from considering an interrelationship in connection with a hierarchical approach. Because of ecological impacts emanating from a company's environment, such as an increased ecologically motivated demand behaviour, an increasing shortage of resources and growing waste disposal and

energy costs as well as the rising average age of the economically active population and the ensuing aggravation of skill shortage, it will also have to be noted that the aspect of sustainability will gain a high significance in the future. In this paper, we have therefore examined existing concepts of sustainability for the various individual stages of hierarchical planning and analysed deficits regarding an integrative concept.

This paper has been structured as follows: Chapter 2 shows the relevance of sustainability. Subsequently, Chapter 3 looks into master production scheduling and shows which social and ecological aspects can be considered at this level of planning. Likewise, batch sizing and resource scheduling are examined in chapters 4 and 5. The concluding Chapter 6 presents our conclusions as well as a summary.

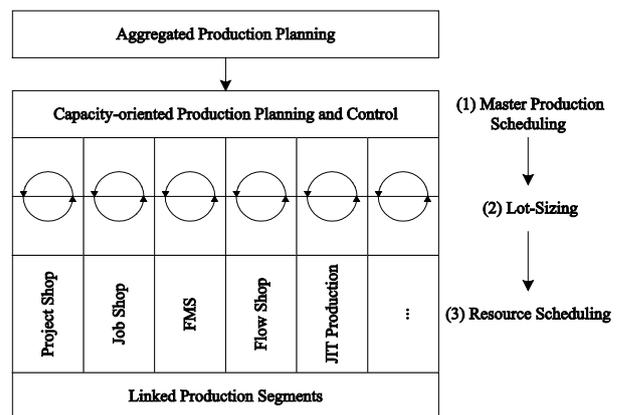


Fig. 1: The concept of hierarchical planning (based on Drexel et al. 1994)

RELEVANCE OF SUSTAINABLE VARIABLES

The idea of sustainability is a frequently discussed notion in science and practice (Andriolo et al. 2014). In the Brundtland report, this notion has been defined as “a development that meets the needs of the present without compromising the ability of future generations to meet

their own needs“ (Brundtland 1987). Any industrial progress should consist of the elements “economy”, “ecology” and “social aspects” which need to form a triad of equivalent priorities since the equal balance of these elements will become a critical task of sustainable corporate governance in the future, particularly due to the aggravation of skill shortage and the rising average age of the economically active population. In industrial companies, this task will first and foremost have to be managed by PPC (Haasis 2008).

Any economic parameters such as storage costs or default charges are traditional parameters in the field of PPC. By considering these parameters, companies succeeded in minimising costs and improving their performance. Besides that, the global intertwinement of markets creates an enormous cost pressure for production companies, forcing them to become more and more flexible; as a result of this, methods of dynamic planning and optimization are required which have to take uncertainties of the market into account and which primarily permit a sufficient scope of action in terms of capacity. (Lanza and Peters 2011).

Due to the ongoing climate changes and the resulting political decisions and requirements, ecological parameters such as efficiency of energy and resources will however also constitute an important dimension in the future. In the framework of the climate and energy package, the EU committed to reduce greenhouse gas emissions by 20 per cent, to increase the percentage of renewable energies to 20 per cent of the total energy demand and to increase energy efficiency by 20 per cent, by the year 2020 in comparison with a development without further efforts to reduce energy consumption. In addition to this, the Paris Agreement concluded during the United Nations Climate Change Conference in 2015 provides to curb global warming to a value significantly below 2 degrees Celcius (if possible 1.5 degrees Celcius). As soon as ecological parameters are taken into account at all levels of hierarchical planning, this will provide an enormous potential for supporting the conformity with these requirements (Müller et al. 2008; Erlach and Westkämper 2009; Vorderwinkler and Heiß 2011). The Fraunhofer IPT estimated that the potential energy savings feasible in Germany in the medium run would range between 25 and 30 per cent (Drescher and Rohde 2009), where mainly cost savings would be a critical criterion, in addition to positive effects on the environment and the climate (Lanza and Peters 2011), a fact which will put even more emphasis on the intertwinement of ecological and economical parameters.

The third dimension to be considered here comprises social parameters which must always be considered as soon as human labour is involved. In spite of different models which are used for work shift planning, job rotation or staff scheduling, the working conditions do

not improve (Schmucker 2014). Among others, the report: “DGB-Index Gute Arbeit 2013” (good work index of the German trade union confederation, 2013) classified the physical stress and work intensity as “poor” and even as “lower medium range” (Schmucker 2014). This report shows that 45 per cent of the work force do not assume to be able to exercise their profession until they will have reached the age of retirement (Schmucker 2014). The consequences of increased stress may be psychosomatic, psychological, or behavioural. As short-term reactions, an increased heart rate (psychosomatic), frustration (psychological), increased failure rate (behavioural & individual) or aggressive behaviour (behavioural & social) can be observed; these factors lead to increased durations of absence from work, resignation as well as psychosomatic diseases (Nerdinger et al. 2014). The general responsibility of companies for their employees urges them to improve working conditions; and the fact that the economically active population decreases, as forecast by the German Federal Ministry of Labour and Social Affairs, increasingly forces companies to boost the performance potential of their employees. A mutual solution to these tasks may for example consist in reducing physical and psychological overstress and in benefiting from learning effects. In addition to the improvements of the work environment and the development of carefully adapted production processes, PPC therefore offers lots of possibilities for improving their employees’ performance potential (Vorderwinkler and Heiß 2011).

MASTER PRODUCTION SCHEDULING

Master production scheduling as a central planning module captures all production segments of a given production site, as well as its main products and its aggregate capacity requirements. The task to be solved consists in preparing production programmes over several time periods and in coordinating these programmes between the various production segments. The starting points of master production scheduling are existing customer orders as well as short-term demand forecasts for end products. The resources needed are organised in groups and units having the same functions and necessitating the same amount of costs (cost centres). The objective is to minimize the relevant costs incurred in connection with production, storage and resources on the basis of the deadlines specified for the target to be reached (Günther and Tempelmeier 2012).

As a rule, the actual capacity needed per unit of quantity to be produced is considered as constant. In case of a purely machine-based production, this assumption is correct. However as soon as manual processes are used in a production system, besides the machine-based processes, the capacity needed for producing a certain unit of quantity of a product will also depend on the employee to whom the job is assigned. However because of the huge complexity and the generally prevailing aggregate approach, a detailed human resources plan-

ning is not expedient in the field of master production scheduling. The objective should rather consist in building up and maintaining a constant performance level of the employees, also against the background of social influences on the production system. This performance level is determined by the respective employee's qualification, experience and workload, however in the framework of master production scheduling, we need to consider employee groups. A possible clustering may for example be set up on the basis of various qualifications.

The nurse-scheduling method for instance considers social effects. In case of a planning horizon ranging between one and three months and a period length of one shift, further parameters are taken into account in addition to the necessity of covering the capacity requirements with the lowest possible number of employees. For instance it is imperative to comply with legal requirements. However any cyclical shift systems which can be set up easily, as suggested by Warner (1976), Warner and Prawda (1972) and Miller et al. (1976), are not sufficient to ensure this. However these models are characterised by a low flexibility towards fluctuating capacity requirements, so that dynamic planning alternatives such as the model created by Smith and Wiggins (1977) have to be given preference (Ozkarahan 1989). There, the individual preferences of employees are additionally taken into account. Human resources scheduling in general has been analysed by literature for a long time. The investigations made by Dantzig (1954) and Edie (1954) were the starting points. The fundamental objective is to cover the required capacities. Here, legal requirements of the Law on Working Hours and stochastic influences such as illness and vacation were also considered. However as a rule, these models assume given capacity requirements which can only be modified by means of advance production or subcontracting. Since the utilisation of learning effects and the reduction of employees' stress exposure have a direct impact on execution times and therefore on the capacity requirements, it is necessary to look at master production scheduling and human resources planning in an integrated manner. Here, it becomes apparent that there are various approaches which are based on a group-related consideration of employees' preferences, simultaneously ensuring the availability of required capacities. However particular approaches considering this explicitly for the PPC and simultaneously including the interdependencies between the employees assigned to the jobs on the one hand and the capacity requirements on the other hand, are still lacking.

Besides capacity requirements, production also generates a certain energy demand which in turn causes further costs. As a rule, the energy price to be considered is assumed to be constant. However in case of particularly energy-intensive (high-consumption) production systems, it may be advantageous to procure energy on the

basis of individual contracts or from the spot market. In the future, we will have to expect fluctuating energy prices. These fluctuations may be of the seasonal or the intra day type. Because of the change-over to regenerative sources of energy, we will have to expect that the amplitude of these fluctuations will keep on rising in the future as witnessed in the past. By planning energy intensive processes in low price periods, companies may benefit from the volatility of energy prices and thus save costs. In addition to variable energy prices, a restriction of energy supplies will have to be considered in the ecological dimension. In the future, it will not be possible to ensure constant energy supplies with the help of regenerative sources of energy (e.g. solar and wind energy) since sufficient energy storage capacities are still lacking (Laux 2013). If phases of low energy generation overlap with peaks of demand, bottle necks will be the result. The idea of integrating these aspects as early as at the moment of the preparation of the master production scheduling has not been considered so far, a fact which constitutes a respective research task.

LOT SIZING

As a result of the previous planning activities, the data of net quantities, as required in the specific periods of time are now available which are used as starting points of lot sizing. These net quantities may be produced on the basis of the "just in time" principle, which will however cause considerable set-up times and therefore set-up costs. Therefore, the task of lot sizing is to combine these required quantities in reasonable batches. The increased storage costs caused here lead to a batch size problem. In addition to this, it is imperative to take existing sequence relations between subordinate and superordinate products and the restricted resource capacities into account, and this will then give rise to a multi-tiered dynamic batch size problem with capacity restrictions. This can be represented in a simplified manner by means of the Multi-Level Capacitated Lot Sizing Problem (MLCLSP). Explanations in this respect may be found in Tempelmeier (2008), Tempelmeier (2012) and Herrmann (2009).

The major part of research work in the field of lot sizing, which is discussed in specialist literature focuses on optimising economic and ecological target parameters. In their economic dimension, the conventional target parameters such as costs of production, ordering, set-up work, storage and transport are looked at in the context of the decision to be taken about lot sizes. In their ecological dimension, a predominant part of research work focuses on different approaches aiming at minimising carbon dioxide emissions (CO₂-emissions) and on the ensuing costs. Here, the interrelation between batch size and CO₂-emissions, resulting from production, transport and storage of this batch size is used as the basis of batch sizing. For example the approaches suggested by Absi et al. (2013) and by Wahab et al. (2011) envisage minimising the costs of batch-size related

CO₂-emissions in connection with transport processes. Other works enlarged this approach even further by integrating the batch size-related CO₂-emission costs of storage (Battini et al. 2014, Bouchery et al. 2012). In case of perishable products, any decisions on batch sizes, which are based on an overestimation of the future product demand may result in the generation of waste. The disposal of these kinds of waste will also generate costs in the form of expenses for CO₂-emissions which need to be taken into account by batch sizing. Approaches in this respect have been provided by the works of Battini et al. (2014) as well as Bonney and Jaber (2011), Arslan and Turkay (2013), Benjaafar et al. (2013), Chen et al. (2013) and Hua et al. (2011) integrated aspects of CO₂-trading (compliance with emission limits, payment of penalties when emission limits are exceeded, prices of emission right trading) into their batch size models and therefore they also aim at minimizing the costs. Besides CO₂ emissions, there are further ecological target parameters which have not been taken into account yet by the research done in the field of batch sizing. As examples, those waste quantities may be referred to here which are brought about by batch packaging (discrepancy between batch size and transport size) or as a result of rejected parts produced in start-up phases after retrofitting processes initiated by batch changes. The energy demand depending on batch sizes in production, storage and transport processes should be included here as well.

The analysis of the state of science regarding the existence of methods which also take the social dimension into account leads us to the research work done by Arslan and Turkay (2013). In this approach, the man hours required for producing, transporting and storing batches are considered as a minimization target; or a limiting value is specified as a side condition of batch sizing. Battini et al. (2014) attributed an indirect significance to the minimisation of transport costs. Since any minimisation of transport costs mostly goes along with a reduction of the number of transports, it is possible to reduce the probability of the occurrence of accidents and traffic jams. The minimisation of emission costs also provides an indirect social contribution, since environmental pollution will thus be reduced and a contribution is made to the protection of the environment for the benefit of the generations to come. A further approach which has a social dimension besides the economic one is the work done by Jaber and Bonney (2007). There, a two-phase model of learning and forgetting is integrated in a classical EMQ-model (Economic manufacture quantity model), where effects of learning and forgetting are taken into account as a function of batch sizes. The two-phase model of learning and forgetting is based on the work submitted by Jaber and Kher (2002), which splits the process of learning and forgetting into a cognitive part and a motor skill part. Further significant social factors which are influenced by the decision on the batch size are in particular the aspects of work ergo-

nomics. The bigger the size of the batch to be manufactured is, the higher the frequency of identical work steps will be which production workers will have to perform repeatedly. This monotony may have a negative impact on the employees' performance potential both in physical and a psychological terms, a fact which will in turn lead to an increase of stress. An approach aiming at illustrating the process of exhaustion was presented by the work of Jaber et al. (2013) which also considered aspects of exhaustion and recovery in addition to a one-phase process of learning and forgetting. However this model was not integrated into a batch size model.

In summary, we need to emphasise that there are different approaches which take the economic, ecological and social dimensions into account, independently to different degrees. However a combination of all these approaches aiming at establishing a really sustainable model has not been achieved yet. Besides that, these approaches partially assume unrestricted capacities, a problem which leads to production schedules that cannot be implemented in entrepreneurial practice. Therefore, an extension will be required here (e.g. including an MLCLSP-model) in addition to a combination of these approaches.

RESOURCE SCHEDULING

Batch sizing is followed by resource scheduling; here, the production orders prepared during batch sizing have to be released on the basis of the previously determined major deadlines and to be allocated to concrete work systems. This elucidates the respective interrelationship between the various planning stages. The length of a period is reduced to any smaller amount of time and any time-consuming processes have to be taken into account (Günther and Tempelmeier 2012). As a possible solution to this problem, we refer to the Resource Constrained Project Scheduling Problem model RCPSP-model presented by Günther and Tempelmeier (2012). The field of application of the RCPSP models is wide and not limited to the original domain of project scheduling (Hartmann and Briskorn 2010). Hartmann and Briskorn (2010) have put various versions of RCPSP together and classified them. Stadler (2005) combined the MLCLSP and the RCPSP and developed an integrated model. However these studies put the focus on the economic dimension.

At a social level, Boysen and Flidner (2011) showed exact and heuristic solutions aiming at reducing the stress exposure of an airport's ground. The workload the ground staff is exposed to depends on the arrival frequency of aeroplanes so that overwork may occur in case of a high arrival frequency of large aeroplanes. Here, the available time window for arrivals is limited by the earliest and the latest times which result from flight distances, speed and quantities of aviation fuels. This may for instance be compared to the planning of production orders. The equivalents of earliest and latest

time are the major deadlines of batch sizing. As a result, the work pressure the employees are exposed to may be reduced by means of an adapted scheduling of the production orders. In addition to a possible reduction of work pressure, Peteghem and Vanhoucke (2015) referred to significant potentials which become available as soon as learning effects are considered regarding the discrete time/resource trade-off scheduling problem (DTRTP). Particularly the reduction of throughput time was presented as a significant result. Furthermore, the authors suggested to consider learning effects in stochastic terms in order to permit an approach that would be more realistic than deterministic methods. In parallel, effects of forgetting must also be considered in addition to learning effects. As regards resource planning, the consequence is that production orders can be planned in such a way that their sequence will generate the highest possible level of learning, that a maximum work load of the employees is not exceeded and that the production targets are achieved according to the time schedule. The available literature offers several approaches to this. However this approach is still lacking a concrete consideration of social parameters for resource scheduling.

In the ecological dimension, energy savings are possible when the modes of operation of machines are considered in resource scheduling (Selmair et al. 2015). Selmair et al. (2015) explained that energy demand and operation time, the latter one being based on resource scheduling, are not directly related which means that energy demand is not directly proportional to the total throughput time, a fact which unveils a considerable potential for optimization. For assessing the energy demand within the optimization process, a flexible energy price has to be chosen, as suggested by Selmair et al. (2015), since as a rule, companies conclude special contracts which stipulate individual energy prices. In addition to this, a possible restriction of the energy supply and of CO₂-emissions has to be considered for the entire planning horizon (for example one day). In a future research task, a sustainable model will have to be developed (e.g. an RCPSP-model) which will aim at a timely achievement of production targets by considering social effects for a more realistic ascertainment of process times and for a reduction of employees' exposure to work pressure as well as ecological effects for a reduction of energy costs.

CONCLUSION

This paper presented various possibilities which could generate an improvement both in the social and in the ecological dimension besides an efficient and timely achievement of production targets, along with a reduction of costs. Here; the tasks of PPC have to be fulfilled by means of hierarchical planning methods, since their application will avoid the disadvantages occurring in a simultaneous planning approach such as a lack of available data. By looking at the three different stages of hierarchical planning, we have pointed out that various

models for solving the subproblems are presented in literature; however as a rule these are limited to the economic dimension.

In the section of master production scheduling we pointed out that a more sustainable approach could help obtaining more realistic results and various kinds of cost savings. In this context, cost savings benefit from fluctuations in energy prices and can also be achieved by reducing the work pressure employees are exposed to, since reduced work pressure contributes to reducing the frequency of work-related diseases, to increasing the motivation and therefore to achieving lower and more constant execution times. In the field of batch sizes, we have shown that some approaches taking the ecological dimension into account do exist. Furthermore it became however apparent that these approaches have not been combined yet in order to create a really sustainable approach. In addition to this, it will be necessary to convert a combined model into a model which is close to reality (e.g. MLCLSP). In resource scheduling, we referred to the RCPSP-model as a multifaceted solution model. In combination with the approaches suggested by Selmair et al. (2015) and Boysen and Fliedner (2011) this will permit reducing work pressure of employees as well as energy costs. Besides that, a significant reduction of throughput times can be achieved as soon as learning effects are taken into account, as demonstrated by Peteghem (2015).

In summary, it can be said that, based on the integration of the ecological and social dimension, hierarchical planning offers supplementary improvement potentials, which may yield additional cost savings and considerably contribute to protecting our resources (in social and ecological terms), besides a timely achievement of the production targets. However the scientific works presented here are exclusively approaches of a sustainable PPC which need to be combined at each level of planning. Besides that, the concrete interdependencies between the various planning stages will have to be analysed explicitly. We also have to point out that the possibilities indicated here are just mere options for the time being. The effects and integration possibilities of these options have to be investigated further.

REFERENCES

- Absi, N.; St. Dauzère-Pérès; S. Kedad-Sidhoum; B. Penz and Ch. Rapine. 2013. "Lot sizing with carbon emission constraints". In *European Journal of Operational Research*, 227 (1), S. 55–61.
- Andriolo, A.; D. Battini; R. W. Grubbström; A. Persona and F. Sgarbossa. 2014. "A century of evolution from Harris's basic lot size model – Survey and research agenda". In *International Journal of Production Economics*, 155 (1), S. 1–23.
- Arslan, M. C. and M. Turkyay. 2013. "EOQ revisited with sustainability considerations". In *Foundations Of Computing And Decision Sciences*, 38 (4), S. 223–249.

- Battini, D.; A. Persona and F. Sgarbossa. 2014. "A sustainable EOQ model – Theoretical formulation and applications". In *International Journal of Production Economics*, 149 (1), S. 145–153.
- Benjaafar, S.; Y. Li and M. Daskin. 2013. "Carbon Footprint and the Management of Supply Chains – Insights from Simple Models". In *Automation Science and Engineering, IEEE Transactions on*, 116 (10), S. 99–116.
- Bonney, M. and M. Y. Jaber. 2011. "Environmentally responsible inventory models – Non-classical models for a non-classical era". In *International Journal of Production Economics*, 133 (1), S. 43–53.
- Bouchery, Y.; A. Ghaffari; Z. Jemai and Y. Dallery. 2012. "Including sustainability criteria into inventory models". In *European Journal of Operational Research*, 222 (2), S. 229–240.
- Boysen, N. and M. Flidner. 2011. "Scheduling aircraft landings to balance workload of ground staff". In *Computers & Industrial Engineering*, 60 (2), S. 206–217.
- Brundtland, G. H. 1987. *Report of the World Commission on environment and development – our common future*. United Nations.
- Chen, X.; S. Benjaafar and A. Elomri. 2013. "The carbon-constrained EOQ". In *Operations Research Letters*, 41 (2), S. 172–179.
- Dantzig, G. B. 1954. Letter to the Editor – A comment on Edie's "Traffic Delays at Toll Booths". In *Journal of the Operations Research Society of America*, 2 (3), S. 339–341.
- Drescher, T. and L. Rohde. 2009. "Energie- und ressourceneffiziente Produktion – Handlungsbedarf für Industrie, Politik und Forschung". In *TOOLS Informationen der Aachener Produktionstechniker*, 2, S. 2–5.
- Drexl, A.; B. Fleischmann; H.-O. Günther; H. Stadtler and H. Tempelmeier. 1994. "Konzeptionelle Grundlagen kapazitätsorientierter PPS-Systeme". *Zeitschrift für betriebswirtschaftliche Forschung*, 46 (12), S. 1022–1045.
- Edie, L. C. 1954. Traffic delays at toll booths. In *Journal of the operations research society of America*, 2 (2), S. 107–138.
- Erlach, K. and E. Westkämper. 2009. *Energiewertstrom – Der Weg zur energieeffizienten Fabrik*. Fraunhofer IRB Verlag, Stuttgart.
- Günther, H.-O. and H. Tempelmeier. 2012. *Produktion und Logistik* (9. Aufl.). Springer, Berlin, Heidelberg, New York.
- Haasis, H.-D. 2008. *Produktions- und Logistikmanagement – Planung und Gestaltung von Wertschöpfungsprozessen* (1. Aufl.). Gabler, Wiesbaden.
- Hartmann, S. and D. Briskorn. 2010. "A survey of variants and extensions of the resource-constrained project scheduling problem". In *European Journal of Operational Research*, 207 (1), S. 1–14.
- Hax, A. C. and H. C. Meal. 1975. "Hierarchical integration of production planning and scheduling". In M. A. Geisler (Hrsg.). *Logistics*. North Holland, Amsterdam, S. 53–69.
- Herrmann, F. 2009. *Logik der Produktionslogistik*. Oldenbourg, München.
- Herrmann, F. and M. Manitz. 2015. "Ein hierarchisches Planungskonzept zur operative Produktionsplanung und steuerung". In Th. Claus; F. Herrmann; and M. Manitz (Hrsg.). *Produktionsplanung und -steuerung*. Springer Gabler, Berlin, Heidelberg, S. 7–22.
- Hua, G.; T. C. E. Cheng and S. Wang. 2011. "Managing carbon footprints in inventory management". In *International Journal of Production Economics*, 132 (2), S. 178–185.
- Jaber, M. Y. and H. V. Kher. 2002. "The dual-phase learning–forgetting model". In *International Journal of Production Economics*, 76 (3), S. 229–242.
- Jaber, Y. M. and M. Bonney. 2007. "Economic manufacture quantity (EMQ) model with lot-size dependent learning and forgetting rates". In *International Journal of Production Economics*, 108 (1), S. 359–367.
- Jaber, M. Y.; Z. S. Givi and W. P. Neumann. 2013. "Incorporating human fatigue and recovery into the learning–forgetting process". In *Applied Mathematical Modelling*, 37 (12), S. 7287–7299.
- Lanza, G. and S. Peters. 2011. "Effizienzsteigerung von Produktionssystemen – Ein Ansatz auf Basis stochastisch dynamischer Optimierung". In *Zeitschrift für Wirtschaftlichen Fabrikbetrieb*, 106 (6), S. 418–422.
- Laux, M. (2013). *Energiewende! Aber wie? Energiespeicher als intelligente Schlüssel für den deutschen Energiemarkt nach dem EnWG, EEG und StromStG*. Bachelor+ Master Publication.
- Miller, H.; W. P. Pierskalla and G. J. Rath. 1976. "Nurse scheduling using mathematical programming". In *Operations Research*, 24 (5), S. 857–870.
- Müller, E.; J. Engelmann and J. Strauch. 2008. *Energieeffizienz als Zielgröße in der Fabrikplanung*. Wt Werkstattplanung online. H 7 / (8), S. 634–639.
- Nerdinger, F. W.; G. Blickle and N. Schaper. 2014. *Arbeits- und Organisationspsychologie* (3. Aufl.). Springer, Berlin, Heidelberg.
- Ozkarahan, I. 1989. "A flexible nurse scheduling support system". In *Computer methods and programs in biomedicine*, 30 (2), S. 145–153.
- Peteghem, V. van and M. Vanhoucke. 2015. "Influence of learning in resource-constrained project scheduling". In *Computers & Industrial Engineering*, 87, S. 569–579.
- Schmucker, R. 2014. *DGB-Index Gute Arbeit – Der Report 2013* (1. Aufl.). Berlin.
- Selmair, M.; F. Herrmann; T. Claus and E. Teich. 2015. "Potentiale in der Reduzierung des Gesamtenergieverbrauchs einer Werkstattfertigung in der Maschinenbelegungsplanung". In M. Rabe and U. Clausen. *Simulation in Production and Logistics*. Fraunhofer IRB Verlag, Stuttgart, S. 177–186.
- Smith, L. D. and A. Wiggins. 1977. "A computer-based nurse scheduling system". In *Computers & Operations Research*, 4 (3), S. 195–212.
- Stadtler, H. 2005. "Multilevel capacitated lot-sizing and resource-constrained project scheduling – An integrating perspective". In *International journal of production research*, 43 (24), S. 5253–5270.
- Tempelmeier, H. 2008. *Material-Logistik – Modelle und Algorithmen für die Produktionsplanung und -steuerung in Advanced-Planning-Systemen* (7. Aufl.). Springer, Berlin, Heidelberg.
- Tempelmeier, H. 2012. *Dynamische Losgrößenplanung in Supply Chains*. Books on Demand, Norderstedt.
- Vorderwinkler, M.; H. Heiß. 2011. *Nachhaltige Produktionsregeln*. Berichte aus Energie- und Umweltforschung 40/2011. Bundesministerium für Verkehr, Innovation und Technologie.
- Wahab, M. I. M.; S. M. H. Mamun and P. Ongkunaruk. 2011. "EOQ models for a coordinated two-level international supply chain considering imperfect and environmental impact". In *International Journal of Production Economics*, 134 (1), S.151–158.
- Warner, D. M. and J. Prawda. 1972. "A mathematical programming model for scheduling nursing personnel in a

hospital". In *Management Science*, 19 (4-part-1), S. 411-422.

Warner, D.M. 1976. "Scheduling nursing personnel according to nurses preferences – a mathematical approach". In *Operations Research*, 24 (5), S. 842-856.

AUTHOR BIOGRAPHIES



MARCO TROST is doctoral student at the Department of Business Science at the Dresden Technical University and he is sponsored by the European Social Fund (ESF). His e-mail address is: Marco.Trost@mailbox.tu-dresden.de.



PROF. DR. THORSTEN CLAUS holds the professorship for Production and Information Technology at the International Institute (IHI) Zittau, a central academic unit of Dresden Technical University and he is the director of the International Institute (IHI) Zittau. His e-mail address is: Thorsten.Claus@tu-dresden.de.



PROF. DR. FRANK HERRMANN holds the professorship for operative production planning and control at the OTH Regensburg and he is the head of the Innovation and Competence Centre for Production Logistics and Factory Planning (IPF). His e-mail address is: Frank.Herrmann@OTH-Regensburg.de.



ENRICO TEICH is doctoral student at the Department of Business Science at the Dresden Technical University and he is research associate at the professorship for Production and Information Technology at the International Institute (IHI) Zittau, a central academic unit of Dresden Technical University. His e-mail address is: Enrico.Teich@tu-dresden.de.



MAXIMILIAN SELMAIR is doctoral student at the Department of Business Science at the Dresden Technical University. His e-mail address is: Maximilian.Selmair@mailbox.tu-dresden.de

A SUPPLY CHAIN OPTIMIZATION FRAMEWORK FOR CO₂ EMISSION REDUCTION: CASE OF THE NETHERLANDS

Narayan Kalyanarengan Ravi^{a*}, Edwin Zondervan^b, Martin Van Sint Annaland^a, J.C. (Jan) Fransoo^c, Johan Grievink^d

a – Department of Chemical Engineering & Chemistry, Eindhoven University of Technology, Netherlands

b – Laboratory of Process System Engineering, University of Bremen, Germany

c – Department of Industrial Engineering, Eindhoven University of Technology, Netherlands

d – Department of Chemical Engineering, Delft University of Technology, Netherlands

KEYWORDS

Carbon Capture, CO₂ reduction, CCS, Optimization, Supply Chain, Mathematical Model.

ABSTRACT

A major challenge for the industrial deployment of a CO₂ emission reduction methodology is to reduce the overall cost and the integration of all the nodes in the supply chain for CO₂ emission reduction. In this work, we develop a mixed integer linear optimization model that selects appropriate sources, capture process, transportation network and CO₂ storage sites and optimize for a minimum overall cost. Initially, we screen the sources and storage options available in the Netherlands at different levels of detail (locations and industrial activities) and present the network of major sources and storage sites at the more detailed level. Results for a case study estimate the overall optimized cost to be €47.8 billion for 25 years of operation and 54 Mtpa reduction of CO₂ emissions (30% of the 2013 levels). This work also identifies the preferred technologies for the CO₂ capture and we discuss the reasons behind it. The foremost outcome of this case study is that capture and compression consumes the majority of the costs and that further optimization or introduction of new efficient technologies for capture can cause a major reduction in the overall costs.

INTRODUCTION

The increasing CO₂ concentration in the atmosphere is directly related to the increase in CO₂ emissions from burning and consumption of fossil fuels, leading to global warming, which is an issue of a great concern today (IPCC 2007). The concentration of CO₂ (396 ppmv) in the atmosphere in 2013 is roughly 40% higher than it was before the industrial revolution, with a growth of approximately 2 ppmv/year in the last ten years (IEA, 2014) and the emission in 2013 is about 56% higher than in 1990. In the Netherlands, the high court has ordered the government to have the emissions cut by at least 20% of the 1990 levels within five years from 2015. The targets for CO₂ reduction by 2030, according to the reports from the Environmental Assessment Agency of the Netherlands and the EU policy, are set at 40% of the 1990 levels, showing a strong commitment to reduce anthropogenic CO₂ emissions.

In the Netherlands, out of the CO₂ emissions totaling 180 Mtpa, which were almost constant over the past few years, approximately 109 Mtpa of CO₂ is emitted by stationary sources from the energy and manufacturing sector (equal to approximately 60% of the total emissions). Efficient use of energy, use of alternative fuels and energy sources, and applying geo-engineering approaches (afforestation and reforestation) can all lead to reduction of CO₂ emissions to the atmosphere (Dennis et al. 2014), but CO₂ capture, transport and sequestration/storage (CCS) has been considered as an important strategy for bulk mitigation of CO₂. According to the International Energy Agency's roadmap, 20% of the total CO₂ emissions should be removed by CCS by year 2050 (Zaman and Lee 2013). The stationary sources provide us with an easier opportunity for bulk reduction in CO₂ emissions nationwide.

The CCS process involves the capture and separation of CO₂ in bulk (from either stack gas or other intermediate gas streams) and then isolating it from the atmosphere through geological sequestration. The nodes of the CCS supply chain problem are the CO₂ source(s), capture process(es), transportation via pipeline(s) and the geological storage sites. A major challenge for the industrial deployment of a CO₂ reduction methodology is to reduce the overall cost and the integration of all nodes of the CO₂ reduction system (Hasan et al. 2014).

In this work, we design a network consisting of sources, a capture system (technologies and materials) and the storage sites to be transported to for the Netherlands. We design the network such that the overall costs for 25 years of operation and 54 Mtpa (30% of the 2013 levels) reduction of CO₂ is minimized. We will also evaluate what the preferred post combustion technologies are. Initially, we develop a Mixed-Integer Linear Optimization (MILP) model for the reduction of CO₂ emissions through CCS. The model is represented a set of constraints and an objective function. Later, for the case study, we first screen the sources at different levels of detail (both for locations and industrial activities) and investigate how the level of detail affects the overall costs in order to select the appropriate required level of detailing. Then, we group the clusters of storage options available in the

*Corresponding Author. *E-mail address* – N.Kalyanarengan.Ravi@student.tue.nl or narayenkr@gmail.com

Netherlands according to the geographical locations to present the network of major sources and storage sites. Having established the supply chain structure, we use the model for minimizing the overall costs in order to find the optimal network connecting sources and storage sites. Finally, we discuss the results and the outcomes obtained and the reasoning behind it.

PROBLEM STATEMENT

The whole network consisting of CO₂ sources, capturing CO₂ from the sources with the technologies and materials available, and transporting it to the storage sites can be viewed as a supply chain network problem (Hasan et al. 2014). Sources can be seen as the suppliers of CO₂, and capacity restrictions for each storage site can be related to the demands of each site which are satisfied by transporting the CO₂ from the capture plant to the storage sites through a pipeline. Basically, the supply chain consists of sources, plants with capture technology and materials and geological storage sites (see Fig. 1). In this work, we have considered that the capture plants are located in the source site to avoid transport of flue gases.

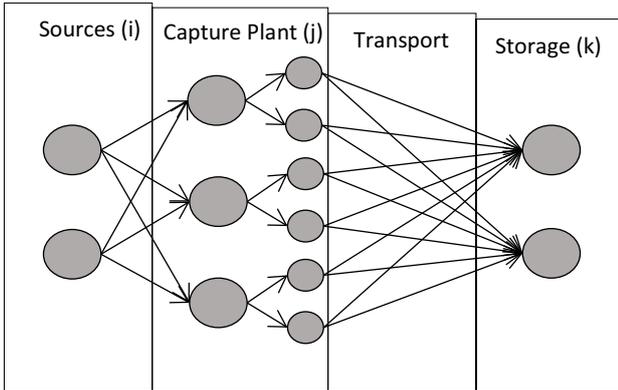


Figure 1 Carbon Capture and Storage Scheme

The problem statement is as follows:

Given:

1. Sources: type & location, yearly CO₂ emissions and compositions
2. CO₂ capture and compression technologies: materials and costs
3. Transportation: distance and quantity to be transported, transportation mode and costs
4. Sequestration/storage: type, location, storage capacity, injection costs and storage limit
5. CO₂ reduction target

Determine:

1. Source and the quantity to be captured
2. Technology and material combination to be used for the CO₂ capture of each selected source
3. Sequestration/storage sites to be used and quantity to be stored in each site
4. Network topology to capture, transport & store CO₂

The objective of the model is to minimize the overall CCS network costs, leading to an optimized structure.

CCS SUPPLY CHAIN NETWORK MODEL DEVELOPMENT

We setup a Mixed-Integer Linear Program (MILP) model to solve the supply chain problem presented in the previous section.

Basic Modelling Assumptions

- The source and capture plants are considered to be in the same and fixed location to avoid transportation of flue gases.
- One to one coupling of source and capture nodes. This means, it is assumed that one source node can be connected to only one capture node and one capture node can receive from only one source node.
- No alternative competing mode of transport to pipeline transport is considered.
- A source node can be connected to only one storage node, but a storage node can receive from multiple source nodes.
- Profit functions such as utilization, carbon tax, etc. are not considered.
- Network structure remains constant throughout the chosen time horizon of 25 years.

MINIMIZE

$$C = \sum_{i,j,k} (CC_{i,j,k} + TC_{i,j,k} + SC_{i,j,k}) \quad (1)$$

s.t.

$$\sum_{j,k} X_{i,j,k} \leq 1 \quad \forall i \in I \quad (2)$$

$$\sum_{i,j} CS_i * FR_{i,j,k} \leq \frac{CU_k^{max}}{Yrs} \quad \forall k \in K \quad (3)$$

$$\sum_{i,j,k} CS_i * FR_{i,j,k} \geq 54 \quad (4)$$

$$FR_{i,j,k} \leq 0.9 * X_{i,j,k} \quad \forall (i,j,k) \in (I,J,K) \quad (5)$$

Eq. 1 shows the objective C , overall costs, as a sum of capture and compression costs ($CC_{i,j,k}$), transportation costs ($TC_{i,j,k}$) and storage costs ($SC_{i,j,k}$). $X_{i,j,k}$ is a binary decision variable that selects a source 'i' and only one suitable technology-material combination 'j' and a storage site 'k' per source and Eq. 2 is a constraint to facilitate this. CS_i is the total emissions from source 'i' and $FR_{i,j,k}$ is 0-1 continuous variable that gives the fraction of CO₂ that is going to be captured from source 'i'. Eq. 3 ensures that the maximum storage limit of each storage site 'k' (CU_k^{max}) is not exceeded. 'Yrs' appearing in the Eq. 3 means the number of years of operation (25 years in our case). Eq. 4 checks if the minimum targeted CO₂ reduction of 54 Mtpa (30% of the 2013 levels) is achieved. Eq. 5 is a constraint, which makes sure that if a source is selected, no more than 90% is captured from that source. The additional computational benefit is avoidance of the

multiplication of variables $FR_{i,j,k}$ and $X_{i,j,k}$ and thereby linearizing the model reported by Hasan et al. (2014). Before going into the details of costs of capture and compression, we need to decide on the technologies and materials to be considered. The four leading capture and compression technologies selected based on maturity and Total Readiness Level are Absorption, Pressure Swing Adsorption (PSA), Vacuum Swing Adsorption (VSA) and Membrane separation (Abanades et al. 2015; Hasan et al. 2014; Zaman and Lee 2013).

$$CC_{i,j,k} = (IC_{i,j,k} + OC_{i,j,k} + DC_{i,j,k}) * Yrs \quad (6)$$

Eq. 6 shows capture and compression costs as a sum of Investment costs ($IC_{i,j,k}$), Operating costs ($OC_{i,j,k}$) and the flue gas Dehydration costs ($DC_{i,j,k}$). Optimizing the capture and compression costs, which depends on flue gas composition and flow rate, is an important step towards reducing the total cost and there have been various efforts to optimize the overall and individual processes. Hasan et al. in their work have optimized various capture and compression technologies and materials and reported the costs for the leading technologies and material combinations in terms of CO_2 composition (X_{CO_2}) and flue gas flow rates (F_i in mol/s) (Hasan et al. 2012a; Hasan et al. 2012b; Hasan et al. 2014). The basic assumptions considered in their cost model are that the technology-material combination is able to capture at least 90% of CO_2 from the flue gas with the least product purity of 90% CO_2 at 150 bar pressure of CO_2 product.

$$IC_{i,j,k} \left(\frac{\text{€}}{\text{yr}} \right) = \alpha * X_{i,j,k} + (\beta x_{CO_2}^n + \gamma) F_i^m \quad (7)$$

$$* (m_{11} FR_{i,j,k} + m_{12} X_{i,j,k})$$

$$OC_{i,j,k} \left(\frac{\text{€}}{\text{yr}} \right) = \alpha_o * X_{i,j,k} \quad (8)$$

$$+ (\beta_o x_{CO_2}^{n_o} + \gamma_o) F_i^{m_o}$$

$$* (m_{21} FR_{i,j,k} + m_{22} X_{i,j,k})$$

Eq. 7 and 8 shows the linearized version for the investment and operating costs per year presented by Hasan et al. (2012; 2014) and the cost model's assumptions and basis can be found in their work. Their model mainly becomes non-linear because of the exponent in $FR_{i,j,k}$. For each of the 13 technology/material combinations considered, the costs are linearized with less than 5% overall relative error compared to the original model. Linearization also allows the model to choose the $FR_{i,j,k}$ freely, rather than assuming it constant as was done by Hasan et al. (2014). The flue gas dehydration costs contribute 9.28 €/t CO_2 captured uniformly. Fig. 2 shows the capture and compression costs as a function of the composition of CO_2 in the flue gas, for a constant flue gas flow rate of 10 kmol/s and $FR_{i,j,k} = 0.9$. The figure is very similar to that provided by Hasan et al. (2014). It can be clearly seen that absorption is preferred for cases with a very low CO_2 composition in the flue gas, whereas adsorption is preferred for cases with higher

compositions. This also shows that the applied linearization does not significantly change the costs of the various material-technology combinations and provides results almost the same as that by the original model presented by Hasan et al. (2014).

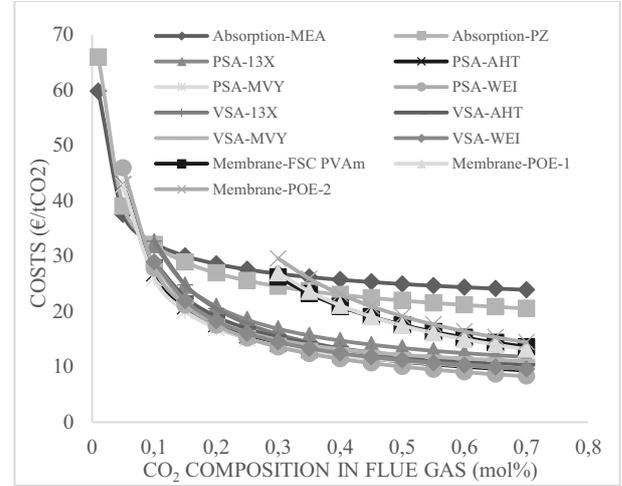


Figure 2 Capture and compression costs for different technology material combinations (Flue gas flow rate = 10 kmol/s)

Modeling of the transportation node(s) also received attention. The review by Knoope et al. (2013) gives a good overview of all the available models. In our work, we use the model presented by Serpa et al. (2011), as it provides us with a linear model and also cost as a function of the quantity transported. We consider a terrain factor, F_T of 1.2, (which can also be taken as a correction factor for distances) and we also add 16 kms to the distance ($D_{i,k}$) for access to a suitable injection site within storage formation (Dahowski et al. 2004). Eq. 9 shows the function for the transport cost that we use in this model. The yearly operation and maintenance costs (OM_t) of transportation are taken as 4% of the investment costs. There are also no distinction made between transportation costs in land and sea.

$$TC_{i,j,k} = Investment + Operating Cost$$

$$= [(\alpha_t * CS_i FR_{i,j,k} + \beta_t * X_{i,j,k}) * F_T * (D_{i,k} + 16)] \quad (9)$$

$$+ OM_t * Investment cost * Yrs$$

For the storage and injection costs, Jansen et al. (2011) give an average investment (I_{well}) and operating costs (OM_{well}) per well and to calculate the number of wells, we use a parameter maximum injection capacity per well (IC^{max}) given by Hasan et al. (2014). Although the well construction, operation and maintenance depend on the type of the storage site and individual well characteristics (like depth, location – offshore & onshore etc.), we assume it to be a constant for the simplicity of the model itself. Eq. 10 shows the storage and injection cost that we use in our model.

$$SC_{i,j,k} = Investment + Operating Cost$$

$$= (I_{well} + Yrs * OM_{well}) \left(\frac{CS_i FR_{i,j,k}}{IC^{max}} \right) \quad (10)$$

CASE STUDY

Data analysis and interpretation

Sources

Data for the CO₂ sources are obtained from the Netherlands Government's Pollutant Release and Transfer Register database and the "Centraal Bureau voor de Statistiek" for the year 2013. The database divides sources with different levels of detailing, according to their location (Total, Province level, Community (municipality) level and Individual location) and industrial activities (Sector level, Sub-Sector level, and Individual Activity level). Initially, to analyze the data, Province – Subsector combination was taken as the others are either less detailed or too detailed. In the total emissions of 180 Mtpa, 242 large stationary sources (leaving out emissions from educational institutions, recreation clubs, etc.) account for ~109 Mtpa, approximately 60% of the total emissions. Out of those 242 sources, the top 35 sources (all ≥ 0.5 Mtpa) account for ~98 Mtpa. We decided to go into different levels of detailing with the same criteria and consider sources only above 0.5 Mtpa emissions.

Four different combinations are considered:

- Province – Sub-Sector
- Province – Individual Activity
- Community – Sub-Sector
- Community – Individual Activity

Obviously, the number of sources and the total emissions are bound to vary when we go into various levels of detail (see Fig. 3) It can be seen that the number of sources are almost constant around 34 and this may be related to the fact that when going into more detail the larger sources getting split into two or more parts. The emissions decrease initially, as expected, and become almost constant at the community level.

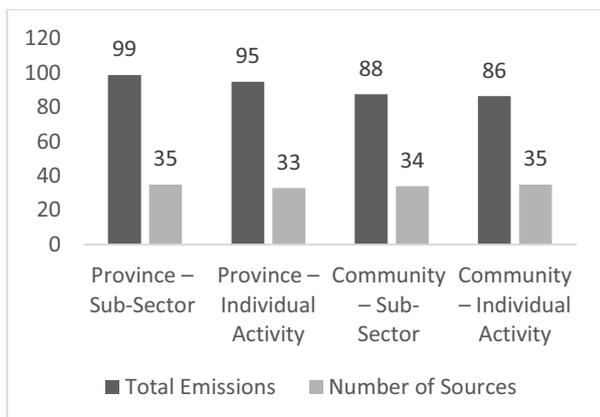


Figure 3 Total Emissions (Mtpa) and Number of sources in each level of detail

Typical CO₂ compositions of flue gas are used for various sources. Fig. 4 shows the composition distribution for the sources at the community-

individual activity level. Most of the sources lie in the composition range of 7% and 20%. Only 3 of the 35 sources have a CO₂ composition above 20%. The flue gas composition plays a major role in the capture costs of CO₂ – the lower the CO₂ composition in the flue gas, the higher the costs.

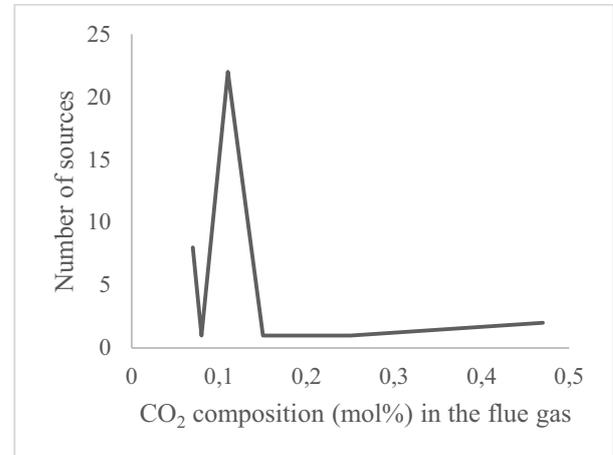


Figure 4 The composition distribution of various sources at the Community-Individual activity level

Fig. 5 shows the objective (total costs for CCS) for different levels of detail. The higher the level of detail, the higher the costs are, as anticipated. It can be noted that the cost becomes almost constant with less than 1% change between the Community-Subsector level and the Community-Individual Activity level. This also shows that going into further detail than the Community-Individual Activity level is not necessary, as there is no noticeable change in the objective.

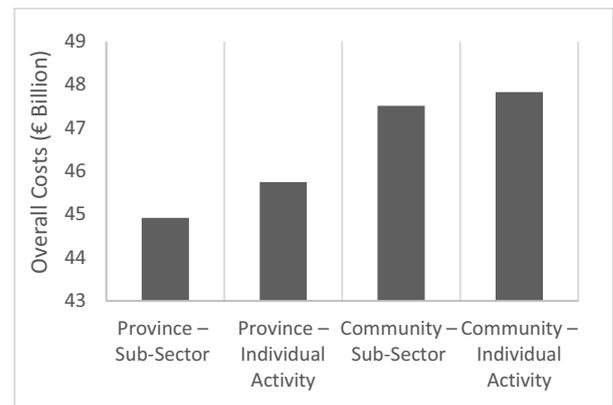


Figure 5 Total costs as a function of the level of detail, clearly showing that the costs become constant at the community - activity level.

Further case & optimization study is evaluated with the data at the level of Community-Individual activity. Fig. 6 shows the location of the sources spread across the Netherlands. It can be clearly seen that the major emitters of CO₂ are located in the western and south-western part of the Netherlands.

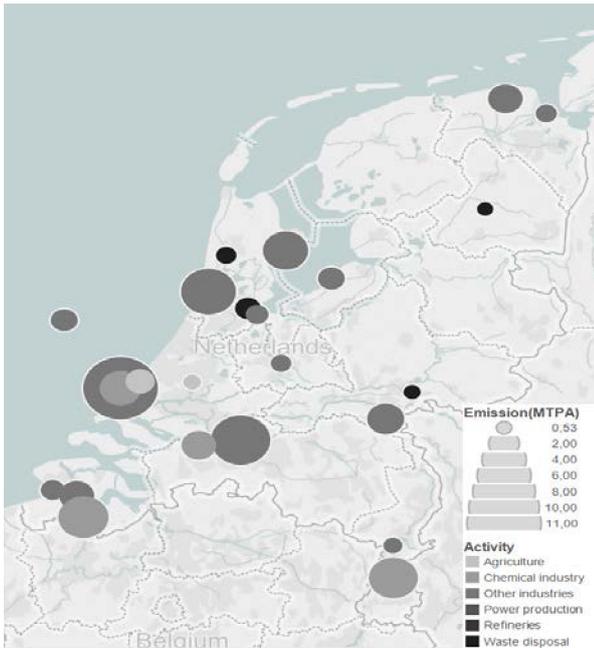


Figure 6. Netherlands map locating the top 35 stationary sources

Storage sites

The storage data were obtained from DHV and TNO (2009), Ramirez et al. (2010), Damen et al. (2009) and Neele et al. (2013). Although there are several hundreds of individual storage sites, the geographical location of the storage sites represented in the above publications on the map of the Netherlands (clusters of storage sites) were grouped manually to reduce the overall problem size to 47 cluster groups, out of which 31 are oil & gas groups, 12 are saline aquifer groups and 4 are groups of coal seams. The storage capacity for each group is estimated on the basis of the known total capacity of each type of storage in the Netherlands. Storage estimation of 47 groups summed to approximately 11 Gt. Out of the 47 storage groups, the top 15 groups contributed to more than 10 Gt of storage and for the ease of implementation, only these 15 storage sites were considered for the case study. Of the 15 storage sites chosen, 11 are oil & gas sites, 3 are saline aquifers and 1 of them is an un-mineable coal seam.

Fig. 7 indicates the geographical location of the grouped storage sites on the map of the Netherlands, where each circle represents the center of the group and size of the circle represents the capacity of the storage. The figure shows that most of the large storage groups are in the north and north-eastern part of the Netherlands. The Groningen site (the biggest circle in Fig. 7) contributes to 7.35 Gt of storage possibility. An important assumption is that all these storage sites are free, ready and available for CO₂ storage and the CO₂ injection platform is going to be built from scratch. Also no costs related to delay by public protests for injection in these storage clusters are assumed.

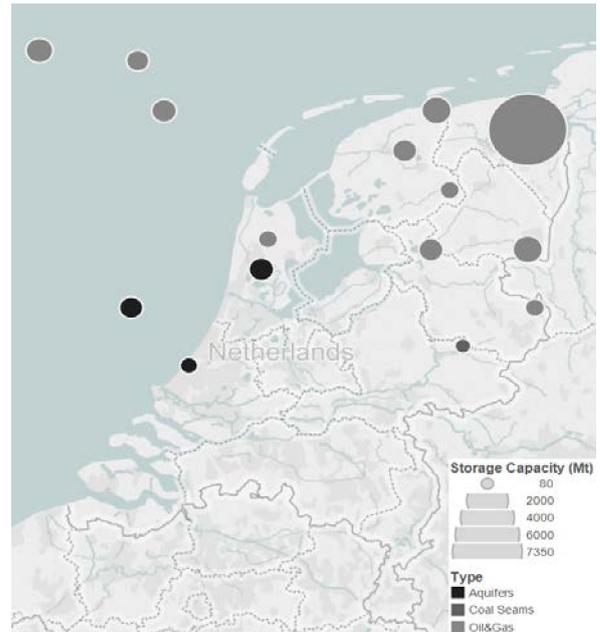


Figure 7. Netherlands map locating the top 15 storage site

Results and discussion

As discussed in the previous section, we consider 35 sources, 13 technology-material combinations for capture & compression and 15 grouped storage sites to inject CO₂. So, the total number of discrete variables are 6825 (35 × 13 × 15). Thus an enormous reduction in the number of sources and storage sites has helped decreasing the size in the model, which also helps in the interpretation of the results. The presented Supply Chain optimization model was used to optimize the costs of the capture of 54 Mtpa of CO₂ and storage for 25 years. A summary of the resulting minimized costs can be found in Table 1.

Table 1 Overall costs and cost per ton basis for the optimal CCS network in the Netherlands

	Overall Costs (€Billion)	Cost (€/tCO ₂ /yr)
Total Expenditure	47.83	35.43
FG Dehydration Costs	12.53	9.28
Capture and compression	30.70	22.74
Sequestration	2.7	2
Transport	1.9	1.42

While, dehydration, storage and transportation add to the total costs, the costs of capture and compression, as expected, is the major contributor. Although we used different cost functions for storage/injection and transportation costs, the cost proportions are very similar to the ones obtained in Hasan et al. (2014). The

total costs for 25 years of operation of CCS is estimated at €47.8 billion and €35.43 per year per ton of CO₂ captured. The storage or injection costs just accounts for 2 €/ton whereas the transportation costs accounts for only 1.42 €/ton. The pipeline costs are often underestimated as the majority of the models reported in the literature keep the cost of natural gas pipelines constructed before 10 – 15 years as the basis, whereas the CO₂ pipelines generally operate at higher pressures (Knoope et al. 2013). The storage costs may also be underrated, but even if the storage costs are 3 or 4 times more, the capture and compression costs with 22.74 €/ton will still remain the largest among all the costs for CO₂ emission reduction. Thus, the main takeaway finding is that the capture processes cause the major lump of expenses and further optimization or invention of new technologies at much lower costs for capture can cause a major change in the overall costs. The optimized network is shown in Fig. 8. The thinner end shows the source and the thicker end shows the storage site and thickness is also proportional to the quantity captured, transported and injected in each storage site. For the optimal design, 18 sources and 9 storage sites are selected by the model. Out of the selected 9 storage sites, 5 storage sites are oil & gas sites, 3 are saline aquifers and 1 of them is an unmineable coal seam.

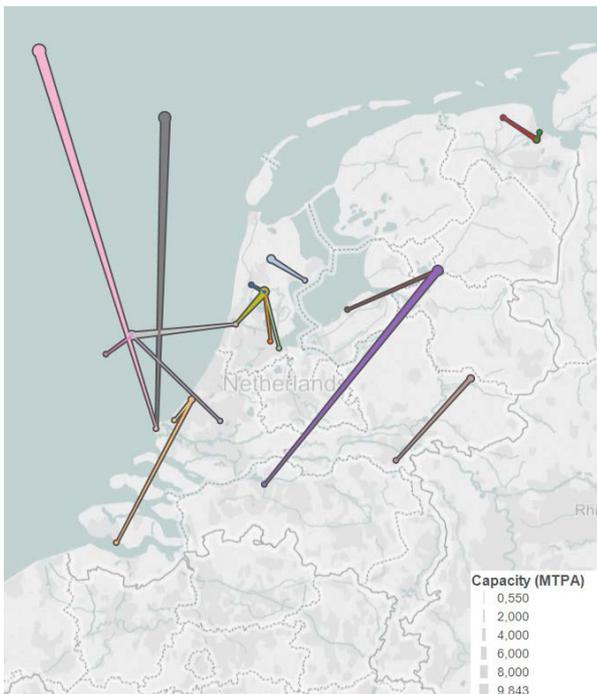


Figure 8 Optimal network for Carbon Capture and Sequestration for the Netherlands

Fig. 9 shows the storage occupancy of each of the storage groups and it can be clearly seen that there still exists more than 85% of the CO₂ storage capacity even after 25 years of operation to reduce 54 Mtpa. The biggest storage site of all, the Groningen gas field (storage site 9 in the Fig. 9), still has almost 100% storage capacity left. To start with, it maybe because of

the straightforward linear relation for costs which doesn't take into account the scale effect of the storage. Furthermore, it is because of the fact that most of the sources selected are from the western or south western part of the Netherlands, whereas the Groningen site is in the Northeastern part of the Netherlands and the transportation cost is comparable to the storage cost.

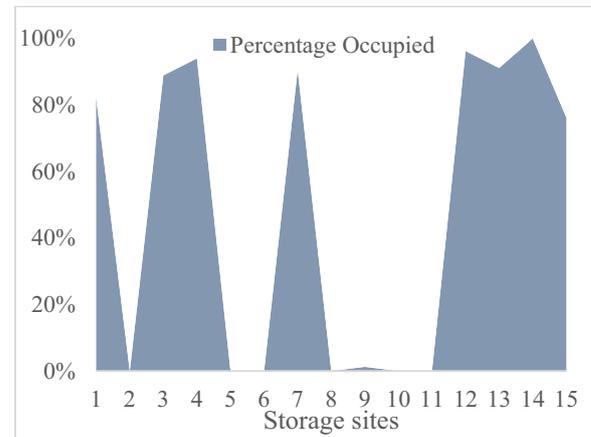


Figure 9 The storage occupancy after the 25 years of optimal operation

In the technology aspect, only 3 out of the 13 technology-material combinations are chosen - 17 of the 18 selected sources use pressure swing adsorption and only 1 use absorption (Fig. 10). In the material feature, MVY (a type of zeolite) based adsorption is strongly preferred over the WEI (another type of zeolite) based one (15 times to two times). In absorption, piperazine (PZ) is preferred over Mono Ethanol Amine (MEA). This shows that the heuristic choice of MEA absorption or absorption in general may not always be the most cost-effective one. Songolzadeh et al. (2014) also found that adsorption is the most preferred post-combustion capture technology at higher feed gas pressures and they also state that adsorption can have a much lower energy consumption and cost for the capture of CO₂. Another reason why adsorption is the most often selected technology in the optimization, is that 17 of the 18 selected sources have a medium to high CO₂ compositions in the feed flue gases (>10%). Absorption is preferred when the concentrations are below 8% at higher flue gas flow rates. This shows that the costs and the selection of the technology depend both on composition and flow rate.

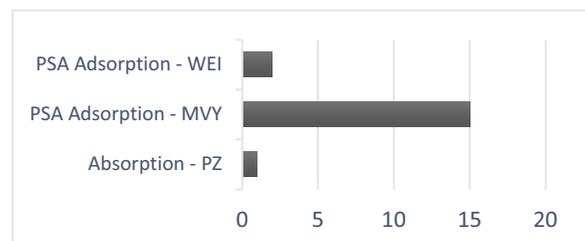


Figure 10 The most preferred technology-material combinations.

CONCLUSIONS & RECOMMENDATIONS

An MILP model is developed and applied to synthesize a national CCS network by optimizing the total costs of this network. Appropriate sources, capture processes, transportation connections and CO₂ storage sites were selected. The MILP model has a linearized relation for the estimation of capture and compression costs. This linearization allows the model to choose the fraction captured from each source instead of assuming it to be a constant. We analyzed different data sets with different detailing for the sources in the Netherlands and came up with a definitive data set, by checking the variation in the objective function, to carry out the case study. The optimal cost achieved by considering the most mature technologies close to commercialization and using an efficient network design, was found to be €47.8 billion for 54 Mtpa of CO₂ reduction in the Netherlands for 25 years of operation. Pressure Swing Adsorption (PSA) was significantly preferred over the heuristic choice of absorption and the difference in costs were also noted to be considerable. It was also concluded that, even after the 25 years of operation, there is still more than 85% of the total storage capacity left across the Netherlands for CO₂ injection. Although the estimate for storage and transportation costs may not be very accurate, a clear conclusion from the relative contribution to the costs is that the capture & compression cost is the major contributor to the total costs. It is therefore recommended to further optimize existing technologies or develop new technologies with much lower capture costs to cause a further major reduction in the overall costs.

REFERENCES

- Abanades, J.C.; B. Arias; A. Lyngfelt; T. Mattisson; D.E. Wiley; H. Li; M.T. Ho; E. Mangano; S. Brandani. Emerging CO₂ capture systems, *International Journal of Greenhouse Gas Control* 40 (2015) 126–166
- Centraal bureau voor de Statistiek, Netherlands
- Dahowski, R.T.; Dooley, J.J.; Davidson, C.L.; Bachu, S.; Gupta, N., 2004. A CO₂ storage supply curve for North America. *PNWD-3471*, 1–92.
- Damen, K.; Andre' Faaij; Wim Turkenburg. Pathways towards large-scale implementation of CO₂ capture and storage: A case study for the Netherlands, *International Greenhouse gas Control* 3 (2009), 217-236
- Dennis, Y. C. Leung; Caramanna, G.; M. Mercedes Maroto-Valer. An overview of current status of carbon-di-oxide capture and storage technologies, *Renewable and Sustainable Energy Reviews*, 39 (2014) 426–443
- Hasan, M. M. F.; Baliban, R. C.; Elia, J. A.; Floudas, C. A., Modeling, simulation, and optimization of postcombustion CO₂ capture for variable feed concentration and flow rate. 1. Chemical absorption and membrane processes. *Ind. Eng. Chem. Res.* 2012, 51, 15642–15664
- Hasan, M. M. F.; Baliban, R. C.; Elia, J. A.; Floudas, C. A., Modeling, simulation, and optimization of post-combustion CO₂ capture for variable feed concentration and flow rate. 2. Pressure swing adsorption and vacuum swing adsorption processes. *Ind. Eng. Chem. Res.* 2012, 51, 15665–15682
- Hasan, M. M. F.; Fani Boukouvala; Eric L. First; and Floudas, C.A., Nationwide, Regional, and Statewide CO₂ Capture, Utilization, and Sequestration Supply Chain Network Optimization, *Ind. Eng. Chem. Res.* 2014, 53, 7489-7506
- IEA, 2014, *CO₂ EMISSIONS FROM FUEL COMBUSTION highlights*, International Energy Agency
- IPCC. 2007. Summary for policymakers. In: climate change 2007: the physical science basis, contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change. *Geneva: World Meteorological Organization /United Nations Environment Program*
- Jansen, F.; Rob Steinz; Boudewijn van Gelder. Potential for CO₂ storage in depleted fields on the Dutch Continental Shelf – Cost estimate for offshore facilities, *Energy Procedia* 4 (2011) 3079–3086
- Knoope, M.M.J.; Ramirez, A.; and Faaij, A.P.C., A state-of-the-art review of techno-economic model predicting the costs of CO₂ pipeline transport, *International Journal of Greenhouse Gas Control* 16 (2013) 241–270
- Neele, Filip; Cor Hofstee; Rob Arts; Vincent Vandeweyer; Manuel Nepveu; Johan ten Veen; Frank Wilschut. Offshore storage options for CO₂ in the Netherlands, *Energy Procedia* 37 (2013) 5220 – 5229
- Pollutant Release and Transfer Register of the Government of the Netherlands
- Ramírez, A.; Saskia Hagedoorn; Leslie Kramers; Ton Wildenborg; Chris Hendriks. Screening CO₂ storage options in The Netherlands, *International Journal of Greenhouse Gas Control* 4 (2010) 367–380
- Serpa, J.; Morbee, J.; Tzimas, E. 2011. Technical and economic characteristics of a CO₂ transmission pipeline infrastructure. *JRC62502*, 1–43
- Songolzadeh, M.; Mansooreh Soleimani; Maryam Takht Ravanchi; and Reza Songolzadeh. Carbon Dioxide Separation from Flue Gases: A Technological Review Emphasizing Reduction in Greenhouse Gas Emissions, *Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, Article ID 828131*
- TNO; DHV 2009 Potential for CO₂ storage in depleted gas fields on the Dutch Continental Shelf
- Zaman, M. and Lee, J. H. Carbon capture from stationary power generation sources: A review of the current status of the technologies, *Korean J. Chem. Eng.*, 30(8), 1497-1526 (2013)

HYBRIDISING LOCAL SEARCH WITH BRANCH-AND-BOUND FOR CONSTRAINED PORTFOLIO SELECTION PROBLEMS

Fang He^{1,2} and Rong Qu¹

¹ The Automated Scheduling, Optimisation and Planning (ASAP) Group, School of Computer Science
The University of Nottingham, Nottingham, NG8 1BB, UK

² Department of Computer Science, Faculty of Science and Technology, University of Westminster,
W1B 2HW, UK

Email: hef@westminster.ac.uk, rxq@cs.nott.ac.uk

KEYWORDS

Hybrid algorithm; Branch-and-Bound; local search; portfolio selection problems

ABSTRACT

In this paper, we investigate a constrained portfolio selection problem with cardinality constraint, minimum size and position constraints, and non-convex transaction cost. A hybrid method named *Local Search Branch-and-Bound* (LS-B&B) which integrates local search with B&B is proposed based on the property of the problem, i.e. cardinality constraint. To eliminate the computational burden which is mainly due to the cardinality constraint, the corresponding set of binary variables is identified as core variables. *Variable fixing* (Bixby, Fenelon et al. 2000) is applied on the core variables, together with a local search, to generate a sequence of simplified sub-problems. The default B&B search then solves these restricted and simplified sub-problems optimally due to their reduced size comparing to the original one. Due to the inherent similar structures in the sub-problems, the solution information is reused to evoke the repairing heuristics and thus accelerate the solving procedure of the sub-problems in B&B. The tight upper bound identified at early stage of the search can discard more sub-problems to speed up the LS-B&B search to the optimal solution to the original problem. Our study is performed on a set of portfolio selection problems with non-convex transaction costs and a number of trading constraints based on the extended mean-variance model. Computational experiments demonstrate the effectiveness of the algorithm by using less computational time.

INTRODUCTION

In this paper, we tackle the single-period portfolio selection problem (PSP). In the problem concerned, a number of transactions can be carried out to adjust the portfolio during a given trading period. We take into account these transaction costs as well as a set of

trading constraints. These include the cardinality constraint (a limit on the total number of assets held in the portfolio, i.e. select k out of n ($k < n$) assets to be held in the portfolio), the minimum position size constraint (bounds on the amount of each asset), the minimum trade size constraint (bounds on the amount of transaction occurred on each asset) and transaction costs. The goal of the problem is to minimize the risk of the adjusted portfolio and the transaction costs incurred, while satisfying the set of trading constraints in feasible portfolios. The aim of this paper is to develop a hybrid method to solve the complex PSP efficiently. The techniques developed here are employed to solve a specific problem, but it could be applied to other variants of PSP with cardinality constraint, and possible other combinatorial problems outside this domain.

If the transaction cost function is linear, then the problem is generally easy to solve. However, a function which better reflects realistic transaction costs is usually non-convex (Konno and Wiyayanayake 2001). Some research show that realistic transaction costs usually include a fixed fee, and thus the cost is relatively higher when the amount of transaction is smaller (Konno and Wiyayanayake 2001, Konno and Wiyayanayake 2002). The transaction cost is thus usually represented by a linear piecewise concave function. This turns the problem into a non-convex optimisation problem, which is more difficult to solve.

In this paper, we propose a new hybrid approach which integrates local search with B&B to solve the non-convex portfolio selection problem *heuristically*. We conceptually divide the decision variables into two parts: the set of core variables which defines the cardinality constraint and the rest of variables. Variable fixing is applied to the core variables. The result of variable fixing has two facets: values (i.e. 0, 1) are assigned to the core binary variables and simplified sub-problem is generated. A local search together with variable fixing are performed on the core variables to generate a sequence of simplified sub-problems. These sub-problems are traversed heuristically to find the

promising sub-problems, i.e. whose lower bounds are not greater than upper bounds. The promising sub-problems then are solved by a default B&B. Value assignments by variable fixing, together with the value assignments by a default B&B, form the complete solutions to the original problem. The best solution to the sub-problem, together with the value assignments by variable fixing approximates an optimal solution to the original problem.

PROBLEM FORMULATION

Consider that an investor is holding an initial portfolio that consists of a set of n assets. To respond to the changes in the market, the investor must review its current portfolio, with the view to carry out a number of transactions. It is assumed that the new portfolio will be held for a fixed time period. The investor's goal is to minimize both the transaction costs occurred and the risk of the assets in the portfolio at the end of the investment period, while satisfying a set of constraints. These constraints typically include meeting the target return, the minimum position size, and the minimum trading size.

Let w_i be the percentage of capital invested in asset i , $i=1, \dots, n$. We shall use a weight vector $w^0 = (w_1^0, w_2^0, \dots, w_n^0)^T$ to denote an initial portfolio. The percentage amount transacted in each asset is specified by weight vector $x = (x_1, x_2, \dots, x_n)^T$, $x_i < 0$ means selling and $x_i > 0$ means buying. A weight vector w denotes the portfolio after the revision. After the transaction, the adjusted portfolio is $w = w^0 + x$, and is held for a fixed period of time. We denote the return of asset i at the end of the investment period as r_i and the expected return of the portfolio as R . We denote the covariance between assets i and j in return as σ_{ij} . We further define $\phi(x)$ as the sum of individual transaction costs associated with each x_i . Based on the basic MV model, the portfolio selection problem with transaction costs can thus be modeled as follows:

$$\min \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \sigma_{ij} w_i w_j + \phi(x) \quad (1)$$

$$s.t. \quad \sum_{i=1}^{i=n} r_i w_i = R \quad (2)$$

$$w = w^0 + x \in F \quad (3)$$

where objective (1) is to minimize the risk of the portfolio and the transaction costs incurred. (2) ensures the expected return. F in (3) represents a set of feasible portfolios subject to all the related constraints. These constraints include the minimum position size, the minimum trading size, etc., which will be detailed next. In this paper, we model the problem as a single

objective problem where (1) is the sum of two objects with the same weights.

The transaction cost is the sum of the transaction costs associated with the assets traded:

$$\phi(x) = \sum_{i=1}^{i=n} \phi_i(x_i)$$

In this paper, we consider a model that includes a fixed fee plus a linear cost, thus leads to a non-convex function, as shown in Fig. 1. This function is also applied in (Lobo, Fazel et al. 2007). The fixed fee charged for buying and selling asset i is denoted as β_i^+ and β_i^- , and the variable costs associated to buying and selling asset i are denoted by α_i^+ and α_i^- . The transaction cost function is given in (4), and shown in Fig. 1:

$$\phi_i(x_i) = \begin{cases} 0, & x_i = 0; \\ \beta_i^+ + \alpha_i^+ x_i, & x_i > 0; \\ \beta_i^- - \alpha_i^- x_i, & x_i < 0; \end{cases} \quad (4)$$

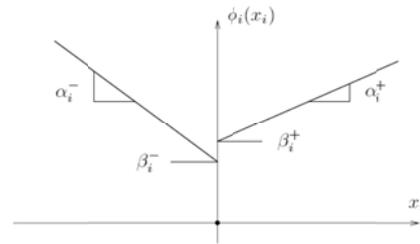


Fig.1 The transaction cost function (Lobo, Fazel et al. 2007)

Problem Model with Transaction Cost and Trading Constraints

Parameter

n	The total number of assets
i	The index of assets, $i=1, \dots, n$
w^0	Initial position of the portfolio
σ_{ij}	Covariance between assets i and j
r_i	Return of asset i at the end of the investment period
R	Expected return of the portfolio
β_i	Fixed cost for buying or selling asset i
α_i	Variable cost rate for buying or selling asset i

w_{\min}	Minimum hold position
x_{\min}	Minimum trading amount
k	Number of assets in the portfolio after transaction

Variable		Feature
w_i	Revised position of the portfolio after transaction	Decision variable
x_i^{buy}	Amount of buying asset i	Decision variable
x_i^{sell}	Amount of selling asset i	Decision variable
z_i	Hold asset i or not in the revised portfolio	Auxiliary variable
z_i^{buy}	Buy asset i or not	Auxiliary variable
z_i^{sell}	Sell asset i or not	Auxiliary variable

There are two groups of variables in the formulation of the problem, as denoted by the “feature” column. w_i , x_i^{buy} , x_i^{sell} are decision variables. z_i , z_i^{buy} and z_i^{sell} are auxiliary variables which are used to formulate the constraints. The column “core variable” denotes which variables are core variables. The selection of the core variables is problem dependent. Several researchers have pointed out that the cardinality constraint presents the greatest computational challenge to the problem (Bienstock 1996, Jobst, Horniman et al. 2001, Stoyan and Kwon 2010, Stoyan and Kwon 2011). Actually, the PSP with cardinality constraint has been recognized to be NP-complete (Bienstock 1996, Mansini and Speranza 1999). To eliminate the cardinality constraint, we identify variables z_i which define the cardinality constraint $\sum_{i=1}^{i=n} z_i = k$ as a set of core variables.

Based on the model PSP, we will introduce two additional reduced models (PSP basic, PSP sub) as follows which will be applied to evaluate the neighbourhood in the local search and to calculate the lower bound:

$$\min \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \sigma_{ij} w_i w_j + \sum_{i=1}^{i=n} \phi_i(x_i) \quad (1) \quad \text{(PSP)}$$

s.t.

$$\sum_{i=1}^{i=n} r_i w_i = R \quad (2)$$

$$w_i = w_i^0 + x_i^{buy} - x_i^{sell}, i = 1, \dots, n \quad (3)$$

$$\sum_{i=1}^{i=n} w_i + \sum_{i=1}^{i=n} \phi_i(x_i) = 1 \quad (5)$$

$$w_i \leq z_i, i = 1, \dots, n \quad (6)$$

$$w_{\min} z_i \leq w_i, i = 1, \dots, n \quad (7)$$

$$\sum_{i=1}^{i=n} z_i = k \quad (8)$$

$$x_{\min} z_i^{buy} \leq x_i^{buy}, i = 1, \dots, n \quad (9)$$

$$x_{\min} z_i^{sell} \leq x_i^{sell}, i = 1, \dots, n \quad (10)$$

$$x_i^{sell} \leq z_i^{sell}, i = 1, \dots, n \quad (11)$$

$$x_i^{buy} \leq z_i^{buy}, i = 1, \dots, n \quad (12)$$

$$z_i^{buy} \leq z_i, i = 1, \dots, n \quad (13)$$

$$z_i^{buy} + z_i^{sell} \leq 1, i = 1, \dots, n \quad (14)$$

$$0 \leq w_i \leq 1, i = 1, \dots, n \quad (15)$$

$$0 \leq x_i^{buy} \leq 1, i = 1, \dots, n \quad (16)$$

$$0 \leq x_i^{sell} \leq 1, i = 1, \dots, n \quad (17)$$

$$z_i, z_i^{buy}, z_i^{sell} \in \{0, 1\}, i = 1, \dots, n \quad (18)$$

$$\min \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \sigma_{ij} w_i w_j \quad (1) \quad \text{(PSP basic)}$$

s.t.

$$\sum_{i=1}^{i=n} r_i w_i = R \quad (2)$$

$$w_i \leq z_i, i = 1, \dots, n \quad (6)$$

$$w_{\min} z_i \leq w_i, i = 1, \dots, n \quad (7)$$

$$\sum_{i=1}^{i=n} z_i = k \quad (8)$$

$$0 \leq w_i \leq 1, i = 1, \dots, n \quad (15)$$

$$z_i \text{ with assignments in } \{0, 1\}, i = 1, \dots, n \quad (18)$$

$$\min \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} \sigma_{ij} w_i w_j + \sum_{i=1}^{i=n} \phi_i(x_i) \quad (1) \quad \text{(PSP sub)}$$

s.t.

$$(2)-(17)$$

$$z_i \text{ with assignments in } \{0, 1\}, i = 1, \dots, n \quad (18)$$

$$z_i^{buy}, z_i^{sell} \in \{0, 1\}, i = 1, \dots, n \quad (19)$$

LS-B&B TO PSP ALGORITHM

In this section, we propose a new hybrid search, named LS-B&B to PSP according to the property of the problem. To the PSP with binary variable z_i we are dealing with, we know that exactly k out of n binary variables will be assigned to 1 in the feasible and optimal solutions. With this knowledge, we can apply variable fixing on a set of variables at one time, resulting into simplified sub-problem. A local search is performed on these set of variables to generate a sequence of sub-problems, and the best solution will be identified among them.

Framework of LS-B&B to PSP

We present the framework of LS- B&B to PSP, as shown in Fig.2.

LS-B&B consists of four main components. The first component is the initialization phase (line 1). In this phase, variable fixing is applied to the core variables to generate a simplified sub-problem. Lower bound and upper bound of the problem are also initialized in this phase.

The second component is a default B&B search (line 7). It is called to solve the sub-problems to optimality. This solution to the sub-problem together with the variable assignments by variable fixing, forms the solution to the original problem.

The third component is a local search (line 9) which is performed on set Z of variable z_i to update sets S and. With the updated S , the sub-problem is updated correspondingly. Therefore, we state that this local search generates a sequence of sub-problems.

The fourth component is an overall search procedure (the while loop). In this search procedure, a local search, variable fixing and a default B&B work together to identify the best solution among the sub-problems by pruning inferior sub-problems and solving the promising sub-problems to optimality.

We present explanations of these components next.

Components of LS-B&B to PSP

Variable fixing

(Hard) variable fixing has been used in MIP context to divide a problem into sub-problems. It assigns values to a subset of variables of the original problem. That is, certain variables are fixed to the given values. Based on the definition of variable fixing in (Bixby, Fenelon et al. 2000, Lazic, Hanafi et al. 2009), we apply this variable fixing to simplify the original problem into sub-problems in the following way. We first denote a subsets S on the binary variable set $B: S \subseteq B$. Then we

fix variables in subsets S to 1, to obtain sub-problems P_{sub_y} as follows:

$$\begin{aligned} P_{sub_y} : \min c^T x \\ s.t. Ax \leq b; \\ x_j = 1, \forall j \in S \subseteq B \neq \emptyset \\ x_j \in [0, 1], \forall j \in C \end{aligned}$$

In this way, we simplify the original problem to a sub-problem. One selection of the subsets S can generate one possible simplified sub-problem of the original problem. Therefore, we apply variable fixing together with a local search to generate a sequence of sub-problems where we will search for the best solution.

LS- B&B
 LB: lower bound;
 UB: upper bound;
 $(h, \mathbf{x}, \mathbf{w}, \mathbf{z})$: a solution $(\mathbf{x}, \mathbf{w}, \mathbf{z})$ of the problem with a corresponding objective value h ;
 solveB&B: a default B&B solver;
 Z : set of z_i ;
 S : subset of Z ;
 P_{org} : the original problem defined by model (PSP);
 P_{sub_y} : sub-problem defined by variable fixing;

- 1: **Initialization phase**
- 2: while (the number of iterations not met)
- 3: If $(LB(P_{sub_y}) \geq UB)$
- 4: prune the sub-problem P_{sub_y} ;
- 5: go to line 9;
- 6: Else
- 7: $(h, \mathbf{x}, \mathbf{w}, \mathbf{z}) = \text{solveB\&B}(P_{sub_y})$;
- 8: if $h < UB$ set $UB = h$;
- 9: perform a **Local search** on set Z ; 10:
 generate sub-problems by variable fixing: $P_{sub_y} =$
 $P_{org} \cup (z_i = 1), z_i \in S$;
- 11: set $(\mathbf{x}^*, \mathbf{w}^*, \mathbf{z}^*)$ as the best solution among all $(\mathbf{x}, \mathbf{w}, \mathbf{z})$
 and h^* be the corresponding objective value;

Fig. 2 The LS-B&B algorithm to PSP

Initialization phase

The main task of the initialization phase is the generation of a sub-problems P_{sub_y} by variable fixing on variables z_i on sets S . From the definition of P_{sub_y} , we can state that P_{sub_y} is P_{org} with the initialization of variables in S to 1.

In the initialization phase, the lower bound is obtained by solving the continuous relaxation of the sub-

problem P_{sub_y} based on model (PSP sub), and the upper bound is set as ∞ .

Default B&B search

As we stated in the framework of LS-B&B, each of the sub-problems itself is still a MIQP problem due to the presence of binary variables z_i^{buy} and z_i^{sell} . However, due to the assignments of variable z_i by variable fixing, the size of the sub-problem is much smaller comparing to the original one. Therefore, sub-problems can be handled by the default B&B. In this paper, the default B&B algorithm in the MIQP solver in CPLEX is applied to solve the promising sub-problems (when $LB(P_{sub_y}) < UB$) to optimality. What is more, the inherent similar structures of the sub-problems enable a very successful reuse of solution information, so the repairing heuristics embedded in solveB&B are evoked to improve the search.

Overall search procedure

The overall search explores the sequence of sub-problems. This is shown in the while loop in Fig.2. In this search, the lower bound of the sub-problem P_{sub_y} is computed by a general QP solver, which relaxes the sub-problem to a continuous problem, i.e. model PSP sub (line 3 in Fig.2). Here, the computation of the lower bound is different from the evaluation of a solution in the local search, which is based on model PSP basic. The objective value of the feasible solution to the concerned sub-problem P_{sub_y} serves as the upper bound of the original problem. If the lower bound of a sub-problem is above the current upper bound found so far, we can discard this sub-problem during the search (line 4 in Fig.2). Otherwise, these promising sub-problems are solved exactly by a default B&B (line 7 in Fig.2). The solutions to the sub-problems together with the assignments of core variables consist of the feasible solutions to the complete original problem. These sub-problems are solved in sequence, and the best solution among them, together with the variable assignments done by variable fixing, approximates the optimal solution to the original problem. The whole procedure terminates by a pre-defined number of iterations in the local search. Therefore, the search is an incomplete search. It cannot guarantee optimality of the solution due to the nature of the local search on core variables z_i .

The local search together with variable fixing creates a sequence of sub-problems which have very similar structures. They only differ in the coefficient or the right-hand side of constraints which are related to z_i . When solving this sequence of sub-problems, the solution information such as the basis list and basis factors from its simplex tableau (i.e., we apply the

extended tableau simplex algorithm in the default MIQP solver) for the current problem are stored, and this can be retrieved and applied to the successive sub-problems. This means the solution information (i.e., basis list and basis factors) of the problem P_{sub_y} can thus be reused to obtain solution to $P_{sub_{y'}}$, so that $P_{sub_{y'}}$ does not need to be solved again from scratch. This solution information reusing thus can evoke the repairing heuristics embedded in the default B&B solver. This solution information reusing has shown to be extremely efficient.

EXPERIMENTAL RESULTS

To evaluate our algorithm on more general benchmark instances, we also concern in this paper the portfolio optimisation instances publicly available in the OR library (ORlibrary), with additional constraints derived from the above real-world problem. Six problem instances are used to test the algorithm proposed in this paper, which can be found at (He and Qu, 2014).

We set the minimum proportion of wealth to be invested in an asset, w_{min} , to 0.01%, and the minimum transaction amount, x_{min} , to 0.01%. We also set the parameters in the transaction cost function α_i to 0.005 and β_i to 0.0001 for all the assets. Other values of k in the cardinality constraint have been tested, ranging from 10 to 150 for different sizes of portfolios.

Evaluations on the LS-B&B algorithm

In LS-B&B, after fixing values for variables z_i by variable fixing and the local search, the resulting MIQP sub-problems are created. If the lower bound of a sub-problem is not greater than the current upper bound (we say it is a promising sub-problem, otherwise it will be pruned), it will be solved by the default B&B in CPLEX12.0. Therefore, when these sub-problems are processed, in conclusion four possible situations could emerge: (1) a sub-problem could be solved by B&B to optimality; (2) the repairing heuristic mechanism imbedded in CPLEX could be evoked and applied to a sub-problem to obtain a feasible solution heuristically; (3) a sub-problem could be pruned; this will happen if the optimal solution under continuous relaxation on model PSP sub is larger than the current upper bound; and (4) the solution of a sub-problem could be infeasible.

Table 1 illustrates the behavior of the above four situations during the processing of sub-problems. The total CPU time of the algorithm is dependent upon the CPU time needed for each situation.

Table 1. Information of sub-problem processing.

Instance	total CPU time	sub-problem solved		sub-problem repaired		sub-problem pruned		sub-problem infeasible	
		Number	Avg CPU time/p	Number	Avg CPU time/p	Number	Avg CPU time/p	Number	Avg CPU time/p
Société Générale	3.16	56	0.01	398	0.006	86	0	60	0
HangS	3.09	184	0.01	178	0.005	120	0	118	0
DAX	9.00	296	0.02	121	0.01	112	0.01	71	0
FTSE	11.44	79	0.08	102	0.025	127	0.02	292	0
S&P	13.55	286	0.04	114	0.01	77	0	123	0
Nikkei	76.97	89	0.40	21	0.36	221	0.08	269	0.06

Table 1 clearly indicates that the CPU time for identifying infeasibility is negligible. The CPU time for pruning the inferior sub-problem is quite efficient. Therefore, the more sub-problems pruned, the more efficient the search is. It can be interpreted from Table 1 that solving sub-problems with repairing heuristics is quite efficient. These repairing heuristics are the results of solution information reuse in the B&B solver. Solving sub-problems exactly is the most time consuming situation comparing with the other three situations.

Comparisons with the default B&B in CPLEX

It is worth noting that LS-B&B is a heuristic approach to the problem. It cannot prove optimality of the solution due to the nature of the local search on core variables z_i , although the sub-problems can be measured by the optimality gap. In order to evaluate the quality of the solutions we obtained from LS-B&B, we compare it against the optimal solution to the problem. It is however very difficult, if not impossible, to obtain and prove the optimal solution to the problems concerned. We therefore calculate the *approximate* optimal solution to the problem concerned by running the default B&B algorithm in CPLEX12.0 for an extensive amount of time.

In the comparison presented in Table 2, we aim to demonstrate the effectiveness of the repairing heuristic evoked in our proposed LS- B&B. Therefore, we present the characteristics of the sub-problems being repaired by heuristic against the characteristics of the default B&B. We compare LS-B&B with the default B&B in Table 2 in terms of the following criteria:

- The number of nodes being processed in B&B to obtain the best integer feasible solution;
- The gap between optimality and the quality of the best feasible solution;

- If the repairing heuristic is evoked and succeed;
- The total CPU time required.

Table 2. Comparisons of default B&B and LS-B&B. + denotes that the repairing heuristics are succeed. All the CPU time is measured in seconds.

		Société Générale	Hang Seng	DAX	FTSE	S&P	Nikkei
Default B&B (original problem)	No. of nodes processed	30	50	150200	147100	130800	35500
	Optimality Gap	0.22%	1.06%	4.66%	3.65%	2.74%	0.44%
	Repair success	No	No	No	No	No	No
	Total CPU time	180					
LS-B&B (sub-problem)	No. of nodes processed	60	80	541800	486700	365800	105000
	Optimality Gap	0.1%	0.29%	4.66%	3.63%	2.74%	0.43%
	Repair success	No	No	No	No	No	No
	Total CPU time	600					
LS-B&B (sub-problem)	No. of nodes processed	0+	0+	50+	30+	30+	50+
	Repair success	Yes	Yes	Yes	Yes	Yes	Yes
	Optimality gap*	0.22%	1.07%	4.65%	3.67%	2.75%	0.44%
	Total CPU time*	3.16	3.09	9.00	11.44	13.55	76.97

In Table 2, in LS-B&B, the number of nodes processed is the average of nodes being processed with repairing heuristics. From Table 2 we can see that by simplifying the problem through variable fixing, the repairing heuristics succeed in LS- B&B approach. The repairing heuristics cannot be evoked by the default B&B while solving the original problem.

Without the simplification, the default B&B needs to explore a much larger number of nodes in the search to obtain feasible solutions, while LS-B&B with simplification requires much less time, shown in Table 2. For example, for the largest instance Nikkei, more than 35,500 nodes have been explored in the default B&B to obtain a feasible solution with a gap of 0.44%.

The optimality gap of solution obtained by LS- B&B is calculated by $gap = (f_{LS} - f_R) / f_R$, where f_{LS} is the objective value obtained by LS- B&B, and f_R is the objective value of continuous relaxation. Table 2 shows that, to achieve solutions of similar quality (as measured by the optimality gap), the CPU time needed by the default B&B is much greater than that required by LS-B&B (e.g. 180 CPU seconds as opposed to 76.97 seconds for the instance Nikkie).

The comparison of LS-B&B with the default B&B can be more clearly illustrated in Fig. 3, which plots of the objective values of LS-B&B and the approximate optimal values obtained by the default B&B with extensive runtime.

It can be seen that LS-B&B converges very well for instances Société Générale, Hang Seng and Nikkei, where the gap between the objective values of LS-B&B and approximate optimal is very small. For instance DAX, the best solution of LS-B&B is even better than the approximate optimal value. For instances FTSE and S&P, the gap is slightly larger. However, it should be noted that LS-B&B spends

significantly less time (3-79 seconds) than the default B&B (180 and 600 seconds).

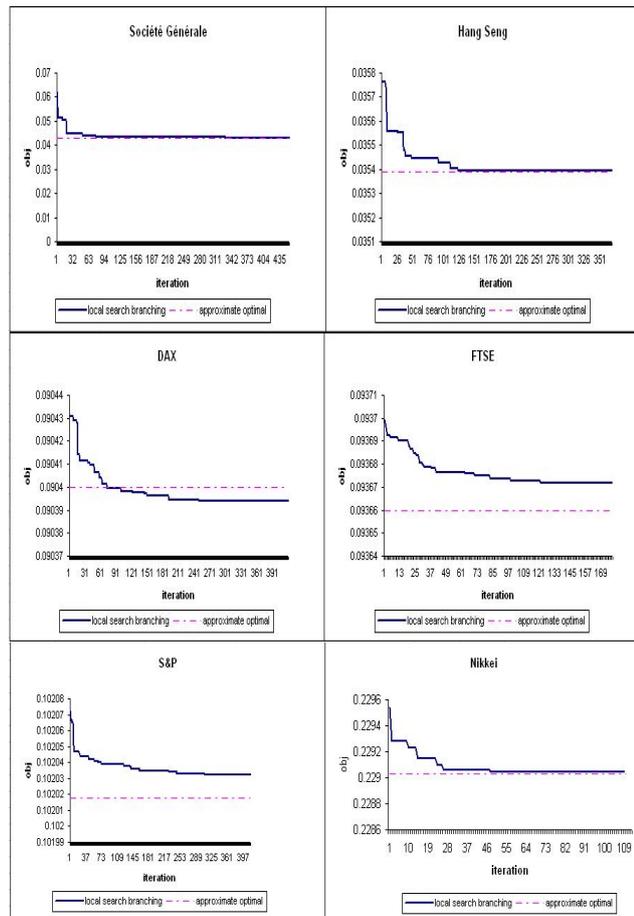


Fig. 3 The gap between LS-B&B and the approximate optimal by the default B&B

CONCLUSIONS

In this paper, we have introduced the hybrid LS-B&B method to solve the portfolio selection problem with practical trading constraints and transaction costs. We have analysed a specific PSP problem which is modelled as MIQP. The hybrid method closely integrates local search with B&B. It implements an incomplete search which aims to seek near optimal solutions in a limited computational time. It simplifies the problem into much smaller sub-problems, which are much easier to solve than the original complete problem, hence can be searched intensively by B&B. It has been demonstrated by our experiments that the repairing heuristics are evoked by solution information reusing in solving sub-problems, thus the successive sub-problems can be solved more efficiently. The heuristic initialization of the core variables in our problem provides a tight upper bound to prune more sub-problems.

REFERENCES

- Bixby, R., M. Felon, Z. Gu, E. Rothberg and R. Wunderling (2000). MIP:Theory and practice--closing the gap. **System Modelling and Optimization: Methods,Theory and Applications 174**: 19-49.
- Bienstock, D. (1996). Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming* 74(2): 121-140.
- Jobst, N. J., M. D. Horniman, C. A. Lucas and G. Mitra (2001). Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints *Quantitative Finance* 1(5): 489-501.
- Hansen, P., N. Mladenovic and D. Urosevic (2001). Variable neighborhood search: Principles and applications. *European Journal of Operational Research* 130(3): 449-467.
- He, F. and R. Qu, A two-stage stochastic mixed-integer program modelling and hybrid solution approach to portfolio selection problems. *Information Sciences*, 289: 190-205, 2014.
- Konno, H. and A. Wiyayanayake (2001). Portfolio optimization problem under concave transaction costs and minimal transaction unit constraints. *Mathematical Programming* 89(2): 233-250.
- Konno, H. and A. Wiyayanayake (2002). Portfolio optimization under D.C. transaction costs and minimal transaction unit constraints. *Journal of Global Optimization* 22(1): 137-154.
- Lazic, J., S. Hanafi, N. Mladenovi and D. Urosevic (2009). Variable neighbourhood decomposition search for 0-1 mixed integer programs. *Computers & Operations Research* 37(6): 1055-1067.
- Mansini, R. and M. G. Speranza (1999). Heuristic algorithms for the portfolio selection problem with minimum transaction lots. *European Journal of Operational Research* 114(2): 219-233.
- Stoyan, S. and R. Kwon (2010). A two-stage stochastic mixed-integer programming approach to the index tracking problem. *Optimization and Engineering* 11(2): 247-275.
- Stoyan, S. J. and R. H. Kwon (2011). A Stochastic-Goal Mixed-Integer Programming approach for integrated stock and bond portfolio optimization. *Comput. Ind. Eng.* 61(4): 1285-1295.

AUTHOR BIOGRAPHIES

Fang He was born in China and obtained her PhD degree from The University of Nottingham, U.K. And she worked as a Research Fellow at the same university for 3 years on modelling and optimisation for combinatorial optimisation problem in real-world applications. This work is conducted during that time. Now she is a lecturer in the Department of Computer Science, University of Westminster.

Rong Qu is an Associate Professor of Computer Science, at the School of Computer Science, The University of Nottingham, U.K. Her research interests are on the modelling and optimisation algorithms (meta-heuristics, mathematical approaches and their hybridisations) to real-world optimisation and scheduling problems.

MODELLING AND OPTIMIZATION OF THE SECOND-HARMONIC RADIATION PATTERN IN DIELECTRIC NANOANTENNAS

Davide Rocco
Luca Carletti
Andrea Locatelli
Costantino De Angelis
University of Brescia
Department of Information Engineering
Via Branze 38, Brescia, 25123, Italy
E-mail: d.rocco003@unibs.it

Valerio F. Gili
Giuseppe Leo
Université Paris Diderot
Matériaux et Phénomènes Quantiques
CNRS, Sorbonne Paris Cité, 10 rue Alice Domon et
Léonie Duquet, F-75013 Paris, France
E-mail: giuseppe.leo@univ-paris-diderot.fr

KEYWORDS

Second Harmonic Generation, dielectric nanoantennas, radiation pattern.

ABSTRACT

We present numerical results that describe how to engineer the radiation pattern of the second harmonic (SH) signal generated by AlGaAs on aluminum oxide all-dielectric nanoantennas. The SH beam divergence is minimized by coherent forward and backward scattering of the radiation emitted at grazing angles from the optical antenna toward a concentric grating grown on the aluminum oxide substrate, whereas the symmetry of the SH mode is converted by introducing a suitably-designed phase shift. The parameters of the structure are optimized through extensive numerical simulations and design guidelines for fabrication are provided.

INTRODUCTION

Optical antennas are a promising research area for their potential application in various areas of nanotechnology: as a matter of fact, their ability to convert propagating radiation into localized subwavelength modes at the nanoscale (Taminiau et al. 2008; Novotny 2008; Koenderink 2009; Devilez et al. 2010; Novotny and van Hulst 2010; Dorfmueller et al. 2011; Miroshnichenko et al. 2011) makes optical aerials highly desirable in many different fields. Antennas have been in common use at radiofrequencies (RF) for more than a century, in a wide variety of applications; as a consequence, well-assessed design rules have been developed in time, and are now available, at RF, for molding the electromagnetic radiation (Balanis 1982). Recent advancements in the fabrication of devices at the nanoscale has allowed to bring many of the concepts of the RF aerials to optics, leading to the development of optical antennas consisting of properly engineered subwavelength metallic and/or dielectric structures (Novotny and van Hulst 2010). In recent years, semiconductor nanoparticles have emerged as a promising alternative to metallic ones for a wide range of nanophotonic applications based on localized resonant modes in the entire visible and near-IR spectral

ranges (Ginn et al. 2012; Albella et al. 2014; Shcherbakov et al. 2014). Different particle geometries can be considered, but mainly spherical and cylindrical antennas have been considered to date. Nanodisks have proven versatile in tailoring of electric and magnetic response, by exploiting two degrees of freedom: radius and height (van de Groep and Polman 2013; Staude et al. 2013). All-dielectric optical antennas also offer unique opportunities for nonlinear optics at the nanoscale with two prominent assets that lack in plasmonics: very low losses leading to high radiation efficiency and multipolar characteristics of both electric and magnetic resonant optical modes potentially leading to the engineering of the radiation pattern (Shcherbakov et al. 2014). It is also worth saying that while in plasmonic nanostructures the optical nonlinear response is dominated by surface nonlinearity (Kauranen and Zayats 2012; Butet et al. 2015; de Ceglia et al. 2015; Celebrano et al. 2015; Finazzi et al. 2007; Dadap 2008), in high-permittivity dielectric nanoantennas the bulk nonlinearity dominates and the properties of the radiation diagram strongly depend on the nonlinear susceptibility elements, the crystallographic orientation, and the input polarization state (Finazzi et al. 2007; Dadap 2008). In this framework, the enhancement of the nonlinear response due to coupling between magnetic and electric dipole resonances has already been observed in third harmonic generation and two-photon absorption (TPA) experiments in Si nanodisks (Shcherbakov et al. 2014, Shcherbakov et al. 2015a; Shcherbakov et al. 2015b). More recently, second harmonic generation (SHG) with an efficiency of 10^{-3} has been theoretically predicted by exploiting the magnetic dipole resonance in AlGaAs structures (Carletti et al. 2015). So far, research in this field has focused on the enhancement of the nonlinear response of nanostructures rather than on the control of the radiation pattern of the nonlinearly generated signals. For example, in (Carletti et al. 2015) a very high SHG efficiency is reported, but the SH signal emitted by the AlGaAs nanodisk has a null along the cylinder axis in the forward and backward directions. Engineering the radiation pattern of SHG for e.g. achieving unidirectional signal emission is of paramount importance for using all-dielectric nanoantennas in applications

such as chemical or biological sensing requiring low power and low cost components, (Albella et al. 2014) because it affects the collection efficiency of experiments in real-life conditions. Although the radiation pattern of SHG in centrosymmetric nanoparticles (Bennemann 1998; Dadap et al. 1999; Mäkitalo et al. 2014; Bautista et al. 2012) and THG in amorphous dielectric particles has been studied, shaping of the SHG emission from non-centrosymmetric nanoparticles exhibiting both magnetic and electric multipole resonances remains almost unexplored. The main goal of this paper is to describe a route based on the structuring of the substrate (Lezec et al. 2002; Garcia-Vidal et al. 2003; Yu et al. 2008; Iwaszczuk et al. 2013) to engineer the radiation pattern of the second harmonic (SH) signal generated by AlGaAs on aluminum oxide all-dielectric nanoantennas. The SH beam divergence is minimized by coherent forward and backward scattering of the radiation emitted at grazing angles from the optical antenna toward a concentric grating structure, whereas the symmetry of the mode is converted by introducing a suitably-designed phase shift. The parameters of the structure are optimized through extensive numerical simulations and design guidelines for fabrication are provided.

SECOND HARMONIC GENERATION IN AlGaAs NANOANTENNAS

In order to demonstrate the control of the radiation profile of the SH field generated by AlGaAs nanoantennas we use frequency-domain simulations implemented using the finite-element-method in COMSOL. The pump beam at the fundamental frequency is assumed to be a plane wave s-polarized along one of the crystalline axes, which are assumed to be aligned with the simulation Cartesian coordinate system axes (see Fig. 1). The second-order nonlinear susceptibility tensor of AlGaAs has only elements of the type $\chi^{(2)}_{ijk}$ with $i \neq j \neq k$ (Carletti et al. 2015). Thus the i -th component of the nonlinear polarization at the SH frequency 2ω is given by:

$$P_i^{(2\omega)} = \varepsilon_0 \chi_{ijk}^{(2)} E_j^{(\omega)} E_k^{(\omega)} \quad (1)$$

where ε_0 is vacuum dielectric constant and $E_j^{(\omega)}$ is the j -th component of the electric field at the pump frequency ω . The nonlinear polarization in Equation (1) is used to define the nonlinear source currents and calculate the SHG from the AlGaAs cylinders. The reference structure that we have considered here is borrowed from (Carletti et al. 2015) and is an $\text{Al}_{0.18}\text{Ga}_{0.82}\text{As}$ cylinder with radius $r = 225$ nm and height $h = 400$ nm on top of an Al_2O_3 substrate.

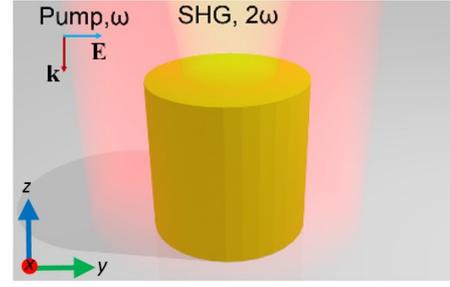


Figure 1: Schematic representation of SHG from a cylinder. The pump beam at ω is a plane wave with wave vector, k , parallel to the cylinder axis.

In order to model the dispersion of the refractive index of $\text{Al}_{0.18}\text{Ga}_{0.82}\text{As}$ we used the analytical model proposed by Gehrsitz (Gehrsitz et al. 2000) which was derived from comparison with measurements. The scattering efficiency (defined as $Q_{sca} = C_{sca}/\pi r^2$ where C_{sca} is the scattering cross-section and r is the cylinder radius) at wavelengths close to the magnetic dipole resonance is reported in Fig. 2(a). It is possible to notice that due to the presence of the substrate the magnetic resonance redshifts ($\lambda = 1655$ nm) with respect to the case of cylinder suspended in air ($\lambda = 1640$ nm).

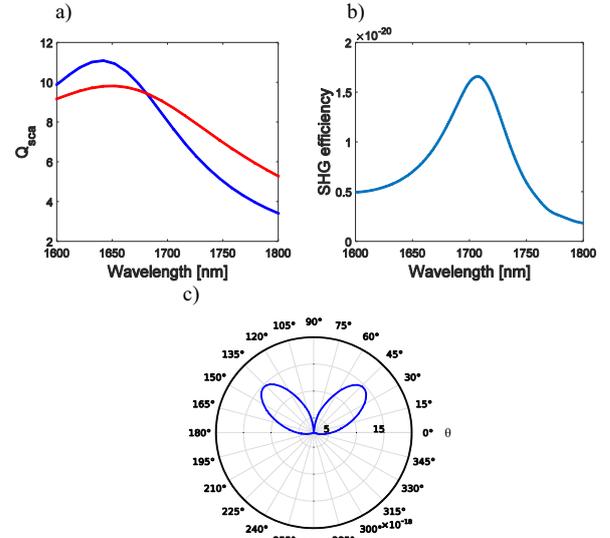


Figure 2: (a) Scattering efficiency Q_{sca} , as a function of wavelength calculated for $r = 225$ nm and $h = 400$ nm, in the case of presence of the substrate (red line) or completely air-surrounded nanostructure (blue line). (b) SHG efficiency as a function of pump wavelength for the same cylinder in the presence of the Al_2O_3 substrate and pump intensity $I_0 = 1.33 \cdot 10^{-3}$ W/m². (c) Far-field radiation pattern of the SH electric field for pumping wavelength $\lambda = 1710$ nm, y - z plane. The incident light is a plane wave with a wave vector, k , parallel to the cylinder axis and the electric field, E , polarized along the y -axis.

We investigate the SHG phenomenon by using the nonlinear polarization induced by the nonlinear susceptibility $\chi^{(2)}$ as a source in COMSOL simulations. We define the SHG efficiency as:

$$\eta_{SHG} = \frac{\int_A \vec{S}_{SH} \cdot \hat{n} da}{I_0 \times \pi r^2} \quad (2)$$

where \vec{S}_{SH} is the Poynting vector of the SH field, \hat{n} is the unit vector normal to the surface A enclosing the antenna and I_0 is the incident field intensity ($I_0 = 1.33 \cdot 10^{-3} \text{ W/m}^2$ in the simulations). We observed that the SHG efficiency peak is for a pumping wavelength of $\lambda = 1710 \text{ nm}$, as reported in Fig. 2(b). Fig. 2(c) shows the far field in the air region at that pumping wavelength. The mode has its twofold symmetry directly transferred to its emission pattern with maximum emission intensity at large off-axis angles; the radiation null in the forward direction ($\theta = 90^\circ$) directly comes from the symmetry of the nanoantenna and the relative orientation between the crystallographic axes and the input polarization state of the pump (Iwaszczuk et al. 2013).

In order to avoid the presence of this null of the radiation diagram one can take advantage of structured pump light beams or resort to different particle geometries; the approach we have considered here explores the use of structuring the substrate similarly to what is done at radio frequency where collimation is achieved through the patterning of the ground plane (Lezec et al. 2002; Garcia-Vidal et al. 2003).

Our beam collimator (described in the next Section) was derived from the plasmonic collimator that was introduced by Yu and Iwaszczuk (Yu et al. 2008; Iwaszczuk et al. 2013) and is essentially based on an interference effect in the near field: as schematically described in Fig. 3, the SHG source (the AlGaAs nonlinear cylinder) couples to the half-ring pattern acting as a 2D ensemble of scatterers that coherently radiate the energy of the SH into the far field. The design of the beam collimator, which is treated in the next section, is thus of utmost importance to shape the multipolar emission pattern of the AlGaAs nanoantenna into a uni-directional beam.

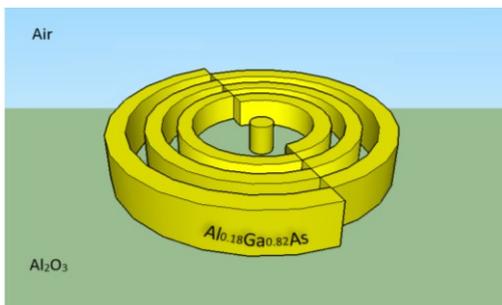


Figure 3: Schematic representation of the AlGaAs concentric grating over the Al_2O_3 substrate.

DESIGN OF THE BEAM COLLIMATOR

In this section we tackle the problem of engineering the radiation pattern of the second harmonic (SH) signal generated by the AlGaAs nanoantennas described in the previous paragraph. We demonstrate that the SH beam divergence can be minimized and the main radiation lobe can be tilted along the cylinder axis by using a concentric grating grown on the aluminum oxide substrate. Our approach mimics the surface plasmon polariton (SPP) assisted scattering from a periodic structure with the groove spacing, width and depth that are optimized for our specific geometry. Indeed, Garcia-Vidal et al. have previously shown that it is possible to collimate and even focus the output of a subwavelength emitter, with an optimized performance for a finite number of grooves in the grating structure (Garcia-Vidal et al. 2003). This principle was later used by Yu (Yu et al. 2008), who fabricated linear and concentric arrays of grooves in the end facet of QCLs for collimation of the output. Agrawal and Nahata used concentric corrugations for resonant enhancement of transmission of THz waves through subwavelength apertures (Agrawal et al. 2005; Agrawal and Nahata 2006) and enhanced coupling between free-space THz beams and wires (Agrawal and Nahata 2007). Fig. 3 illustrates the basic design considered here. In our case we place N concentric corrugations with radial distance d , width a , and depth t_g around the optical antenna. The first step is to design an effective grating for the SH wavelength. We start from a 2D model of the grating. Fig. 4(a) shows the scheme of such a structure that consists of teeth composed of high (AlGaAs) and low (air) index material deposited onto a low-index layer (Al_2O_3). The design parameters for the structure include the grating period (Λ), the grating thickness (t_g), and the duty cycle (DC). The DC is defined as the ratio of the width of the high index material with respect to Λ (Finazzi et al. 2007). Here we assume an infinite thickness of the low-index layer under the grating (t_L). We fix t_g equal to the height of the cylinder. By using the reciprocity principle, we excite the structure with a Gaussian beam at a wavelength of 855 nm (i.e. the wavelength at which the SHG efficiency should be maximum) at normal incidence and with the electric field polarized along the y-axis, and we measure the power that is scattered from one side of the grating towards a numerical probe (depicted by the red line in Fig. 4(a)). We vary the DC and the period of the grating in order to find the values that maximize the lateral scattered power. In our simulations the probe is about 1200 nm far from the grating. We repeated the simulation for different distances of the probe but we found that the result is independent from the distance.

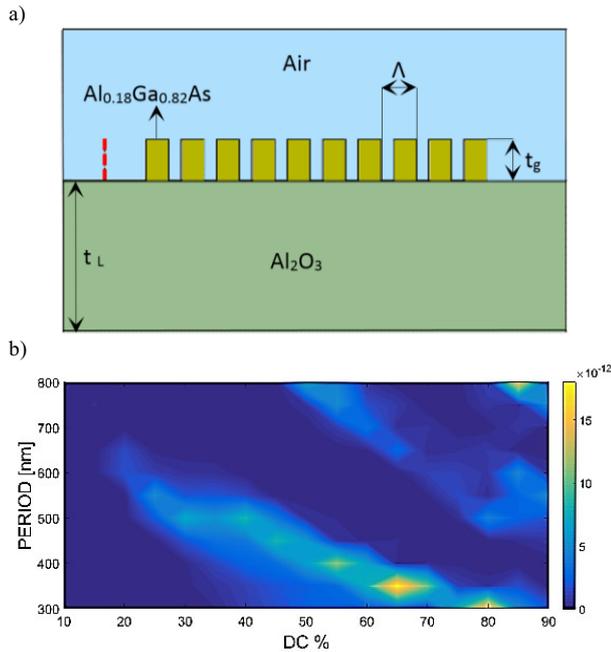


Figure 4: (a) Schematic picture of the system under study: one-dimensional high contrast grating realized in AlGaAs over an Al_2O_3 low-index layer. (b) Scattered power as a function of grating period Λ and duty-cycle for a probe placed 1200 nm far away from the grating.

We discovered that the values that maximize power are a grating period $\Lambda = 350$ nm with a DC equal to 65% that corresponds to a width of the high index material of 227.5 nm. The next step of our design procedure consists in evaluating the optical response of the nanoantenna surrounded by the grating by using a three-dimensional model. In particular, we placed N concentric grooves around the cylinder at a radial distance $d = 380$ nm. That distance is chosen such that it does not affect the linear scattering behaviour of the cylinder at the fundamental frequency. The concentric grating structure will lead to a vertical redirection of the radiation emitted in the xy -plane; however, the quadrupole symmetry of the SHG mode still prohibits that radiation is coupled onto the normal propagation direction. Therefore, as schematically illustrated in Fig. 5, we must introduce a phase shift between the light scattered from each of the two half-rings of the concentric grating structure (see Fig. 5(a)). This will lead to a conversion of the quadrupole mode to a dipolar mode, and thus the far-field pattern will be coupled onto the propagation direction normal to the surface (Dadap 2008). Looking at the magnitude of the far-field in the forward propagation direction ($\theta = 90^\circ$), we notice that the optimal shift between the two half rings is $s = 250$ nm, see Fig. 5(c).

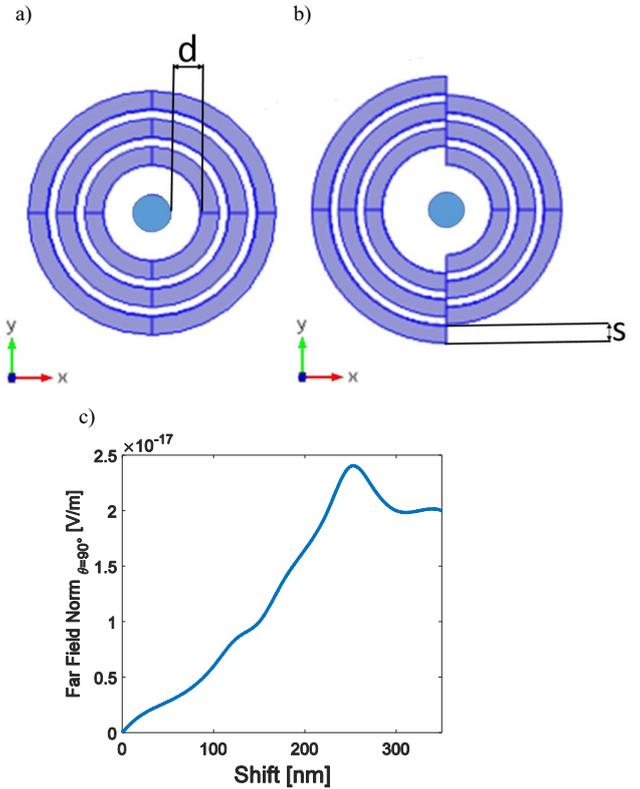


Figure 5: (a) Schematic illustration of a concentric grating structure with radial distance $d = 500$ nm; top view. (b) Concentric grating with phase-shift $s = 250$ nm. (c) Magnitude of the far-field in the forward direction ($\theta = 90^\circ$) as a function of the phase shift s .

The effect of the grating can be verified by looking at the electric field above the cylinder. As it can be seen from Fig. 6(b), when the optimized grating is present the zero at the normal direction ($\theta = 90^\circ$) disappears. Fig. 6(c) shows the calculated far field intensity pattern emitted from the structure with the $N = 3$ grating. We observe a clear collimation of the generated SH field into a narrow, forward-propagating lobe. The asymmetry observed in the emission in the xz plane from the optimized structure is due to the asymmetry introduced in the grating structure in that plane, combined with the usage of a grating with finite number of grooves. We have observed (results not shown here) that, by increasing the number of grooves, this asymmetry is significantly reduced. However, we point out that our interest is in the realization of ultracompact frequency converters operating at the nanoscale, thus we aim at finding a good trade-off between size and performance. Moreover, the simulation of structures with high N is an extremely demanding task and therefore, for the present analysis, we limit ourselves to small values of N .

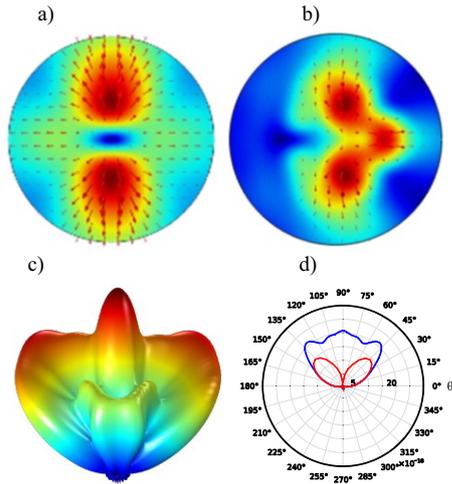


Figure 6: Magnitude of the SH electric field in the x - y plane (top view) in the case of, (a), single cylinder, and, (b), $N = 3$ grating. (c) SH far field in the y - z plane in the case of $N = 3$ grating. (d) Comparison of SH radiation pattern in the air region for pumping wavelength $\lambda = 1710$ nm in the case of single cylinder (red line) and $N = 3$ grating (blue line), y - z plane.

The Second Harmonic Generation efficiency as a function of the pump wavelength for the structure composed of the nanoantenna surrounded by the grating is shown in Fig. 7(a). We can observe a maximum at $\lambda = 1655$ nm, which corresponds to the magnetic dipole resonance of the cylinder alone (see Fig. 2(a)). This may be due to the fact that the structure formed by the cylinder with concentric grating around it has a strong resonance at that wavelength. In addition, the efficiency when the grating is present is slightly higher with respect to the case of the isolated cylinder structure.

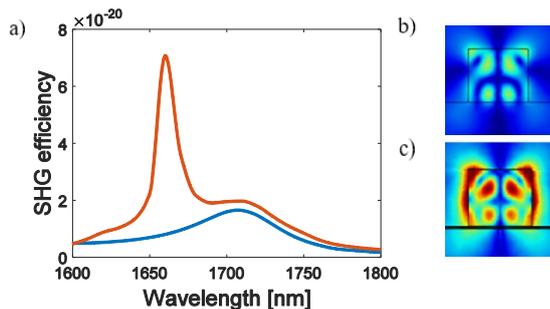


Figure 7: (a) Comparison of SHG efficiency in the case of single cylinder (blue line) and $N = 3$ grating (orange line). (b) Normalized $|E^{SH}|$ on a cross section in the y - z plane at the centre of the cylinder without grating at pumping wavelength $\lambda = 1710$ nm. (c) Normalized $|E^{SH}|$ on a cross section in the y - z plane at the centre of the cylinder with $N = 3$ grating at pumping wavelength $\lambda = 1710$ nm.

CONCLUSION

We report the design of an AlGaAs nanoantenna based on a nanodisk with a concentric grating structure that guarantees a desired SH radiation pattern. We show that, by careful engineering of the surrounding grooves, it is possible to convert a radiation pattern with azimuthal mode number $m = 2$ into a dipolar-like emission profile. This was accomplished by introducing a concentric grating structure that scatters the SH radiation in the forward direction. The symmetry of the generated SH mode was broken by an engineered asymmetry of the grating structure, leading to a highly collimated output.

REFERENCES

- Albella, P., Alcazar de la Osa, R., Moreno, F. and Maier, S. A. 2014. "Electric and magnetic field enhancement with ultralow heat radiation dielectric nanoantennas: consideration for surface-enhanced spectroscopies." *ACS Photonics*, Vol.1, 524-529.
- Agrawal, A., Cao, H., and Nahata, A. 2005. "Time-domain radiative analysis of enhanced transmission through a single subwavelength aperture". *Opt. Express*, Vol.13, 3535-3542.
- Agrawal, A. and Nahata, A. 2006. "Time-domain radiative properties of a single subwavelength aperture surrounded by an exit side surface corrugation." *Opt. Express*, Vol.14, 1973-1981.
- Agrawal, A. and Nahata, A. 2007. "Coupling terahertz radiation onto a metal wire using a subwavelength coaxial aperture." *Opt. Express*, Vol.15, 9022-9028.
- Balanis, C. 1982. *Antenna Theory: Analysis and Design*. Brisbane: J. Wiley, N. J.
- Bautista, G., Huttunen, M. J., Mäkitalo, J., Kontio, J. M., Simonen, J., and Kauranen, M. 2012. "Second-harmonic generation imaging of metal nano-objects with cylindrical vector beams." *Nano Lett.*, Vol.12, 3207-3212.
- Bennemann, K. 1998. *Non-linear Optics in Metals*. Clarendon Press.
- Butet, J., Brevet, P. F., and Martin, O. J. F. 2015. "Optical Second Harmonic Generation in Plasmonic Nanostructure: from Fundamental Principles to Advanced Application." *ACS Nano*, Vol.9, 10545-10562.
- Carletti, L., Locatelli, A., Stepanenko, O., Leo, G. and De Angelis, C. 2015. "Enhanced second-harmonic generation from magnetic resonance in AlGaAs nanoantennas." *Opt. Express*, Vol.23, 26544-26550.
- Celebrano, M., Wu, X., Baselli, M., Großmann, S., Biagioni, P., Locatelli, A., De Angelis, C., Cerullo, G., Osellame, R., Hecht, B., Duò, L., Ciccacci, F. and Finazzi, M. 2015. "Mode matching in multiresonant plasmonic nanoantennas for enhanced second harmonic generation." *Nat. Nanotechnology*, Vol.10, 412-417.
- Dapap, J. I., Shan, J., Eisenthal, K. B. and Heinz, T. F. 1999. "Second-Harmonic Rayleigh Scattering from a Sphere of Centrosymmetric Material." *Phys. Rev. Lett.*, vol.83, 4045-4048.
- Dapap, J. I. 2008. "Optical second-harmonic scattering from cylindrical particles." *Phys. Rev. B*, Vol.78, 205322.
- De Ceglia, D., Vincenti, M. A., De Angelis, C., Locatelli, A., Haus, J. W., and Scalora, M. 2015. "Role of antenna modes and field enhancement in second harmonic

- generation from dipole nanoantennas." *Opt. Express*, Vol.23, 1715.
- Devilez, A., Stout, B. and Bonod, N. 2010. "Compact Metallo-Dielectric Optical Antenna for Ultra Directional and Enhanced Radiative Emission." *ACS Nano*, Vol.4, 3390-3396.
- Dorfmueller, J., Dregely, D., Esslinger, M., Khunsin, W., Vogelgesang, R., Kern, K. and Harald Giessen. 2011. "Near-Field Dynamics of Optical Yagi-Uda Nanoantennas." *Nano Lett.*, Vol.11, 2819-2824.
- Finazzi, M., Biagioni, P., Celebrano, M., and Duò, L. 2007. "Selection rules for second-harmonic generation in nanoparticles." *Phys. Rev. B*, Vol.76, 125414.
- Garcia-Vidal, F. J., Martin-Moreno, L., Lezec, H. J. And Ebbesen, T. W. 2003. "Focusing light with a single subwavelength aperture flanked by surface corrugation." *Appl. Phys. Lett.*, Vol.83, 4500-4502.
- Gehrsitz, S., Reinhart, F. K., Gourgon, C., Herres, N., Vonlanthen, A., and Sigg, H. 2000. "The refractive index of AlxGa1-xAs below the band gap: accurate determination and empirical modelling." *J. App. Phys.*, Vol.87, 7825-7837.
- Ginn, J. C., Brener, I., Peters, D. W., Wendt, J. R., Stevens, J. O., Hines, P. F., Basilio, L. I., Warne, L. K., Inhlfeld, J. F., Clem, P. G., and Sinclair, M. B. 2012. "Realizing optical magnetism from dielectric metamaterials." *Phys. Rev. Lett.*, Vol.108, 097402.
- Iwaszczuk, K., Bisgaard, C. Z., Andronico, A., Leo, G. and Jepsen, P.U. 2013. "Numerical Investigation of Terahertz Emission Properties of Microring Difference-Frequency Resonators." *IEEE Transactions on Terahertz Science and Technology*, Vol.3, 192-199.
- Kauranen, M. and Zayats, A. V. 2012. "Nonlinear Plasmonics." *Nat. Photonics*, Vol.6, 737-748.
- Koenderink, A. F. 2009. "Plasmon Nanoparticle Array Waveguides for Single Photon and Single Plasmon Sources" *Nano Lett.*, Vol.9, 4228-4233.
- Lezec, H. J., Degiron, A., Devaux, E., Linke, R. A., Martin-Moreno, L., Garcia-Vidal, F. J. and Ebbesen, T. W. 2002. "Beaming light from a subwavelength aperture." *Science*, Vol.297, 820-822.
- Makeev, E. V. and Skipetrov, S. E. 2003. "Second harmonic generation in suspensions of spherical particles." *Opt. Commun*, Vol. 224, 139-147.
- Mäkitalo, J., Suuriniemi, S., and Kauranen, M. 2014. "Enforcing symmetries in boundary element formulation of plasmonic and second-harmonic scattering problems." *J. Opt. Soc. Am. A*, Vol.31, 2821-2832.
- Miroshnichenko, A. E., Maksymov, I. S., Davoyan, A. R., Simovski, C., Belov, P., and Kivshar, Y. S. 2011. "An arrayed nanoantenna for broadband light emission and detection." *Phys. Status Solidi RRL*, Vol.5, 347-349.
- Novotny, L. 2008. "Optical antennas tuned to pitch." *Nature (London)*, Vol.455, 887.
- Novotny, L. and van Hulst, N. 2010. "Antennas for light." *Nat. Photonics*, Vol.5, 83-90.
- Shcherbakov, M. R., Neshev, D. N., Hopkins, B., Shorokhov, A. S., Staude, I., Melik-Gaykazyan, E. V., Decker, M., Ezhov, A. A., Miroshnichenko, A. E., Brener, I., Fedyanin, A. A., and Kivshar, Y. S. 2014. "Enhanced third-harmonic generation in silicon nanoparticles driven by magnetic response." *Nano Lett.*, Vol.14, 6488-6492.
- Shcherbakov, M. R., Shorokhov, A., Neshev, D. N., Hopkins, B., Staude, I., Melik-Gaykazyan, E. V., Ezhov, A. A., Miroshnichenko, A. E., Brener, I., Fedyanin, A. A., and Kivshar, Y. S. 2015. "Nonlinear Interference and Tailorable Third-Harmonic Generation from Dielectric Oligomers." *ACS Photonics*, Vol.2, 578-582.
- Shcherbakov, M.R., Vabishchevich, P. P., Shorokhov, A. S., Chong, K.E., Choi, D. Y., Staude, I., Miroshnichenko, A. E., Neshev, D. N., Fedyanin, A. A. and Kivshar, Y. S. 2015. "Ultrafast All-Optical Switching with Magnetic Resonances in Nonlinear Dielectric Nanostructures." *Nano Lett.*, Vol.15, 6985-6990.
- Staude, I., Miroshnichenko, A. E., Decker, M., Fofang, N. T., Liu, S., Gonzales, E., Dominguez, J., Luk, T. S., Neshev, D. N., Brener, I., and Kivshar, Y. S. 2013. "Tailoring directional scattering through magnetic and electric resonances in subwavelength silicon nanodisks." *ACS Nano*, Vol.7, 7824-7832.
- Taminiau, T. H., Stefani, F. D., and van Hulst, N. F. 2008. "Enhanced directional excitation and emission of single emitters by a nano-optical Yagi-Uda antenna." *Opt. Express*, Vol.16, 10858-10866.
- Van de Groep, J., and Polman, A. 2013. "Designing dielectric resonators on substrates: combining magnetic and electric resonances." *Opt. Express*, Vol.21, 26285-26302.
- Yu, N., Fan, J., Wang, Q. J., Pflügl, C., Diehl, L., Edamura, T., Yamanishi, M., Kan, H. and Capasso, F. 2008. "Small-divergence semiconductor lasers by plasmonic collimation." *Nat. Photon.*, Vol.2, 564-570.

AUTHOR BIOGRAPHIES



DAVIDE ROCCO was born in Brescia (Italy) on November 24th, 1989 and went to the University of Brescia, where he studied Electronic and Telecommunication Engineering and took his bachelor degree in May 2012. After that he started the Master Education. He received the Master degree in Communication Technologies and Multimedia at the University of Brescia in 2015. He now attends PhD. study in the Department of Information Engineering, at the University of Brescia. His research interests focus on dielectric optical nanoantennas.



LUCA CARLETTI received the B.Sc. in Information Engineering from University of Padova, Italy in 2008, the M.Sc. in Telecommunication engineering from University of Padova, Italy and the M.Sc. degree in Physics and Nanotechnology from Technical University of Denmark (DTU), Lyngby, Denmark in 2011. In 2014 he received his PhD degree from the Ecole Centrale of Lyon, France working on nonlinear optical phenomena in integrated photonic structures from near to mid infrared wavelength focused on applications in telecommunications and sensing. He is currently working as a research associate at the University of Brescia. His current research activity focuses on the linear and nonlinear optical response of nanoantennas. Dr. Carletti was awarded a TIME double degree scholarship in 2009.



ANDREA LOCATELLI was born in Seriate (Bergamo) on July 3rd, 1977. He received the Master degree (cum laude) in Electronic Engineering and the Ph.D. degree in Information Engineering from the University of Brescia in 2001 and 2005, respectively. Since 2002 he has been carrying out his research activity at the Department of Information Engineering of the University of Brescia. In 2008 he became Assistant Professor in the field of Electrical Engineering (ING-IND/31 Elettrotecnica). He has authored or co-authored more than 120 papers published in international journals, conference proceedings and books. He is a member of IEEE (Photonics Society and Power & Energy Society), Optical Society of America (OSA), European Microwave Association (EuMA), Consorzio Nazionale Interuniversitario per le Scienze Fisiche della Materia (CNISM), Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), and Istituto Nazionale di Ottica (INO). He regularly serve as referee for several international journals, and for national research projects funded by the Italian Ministry for Education (MIUR).



COSTANTINO DE ANGELIS was born in Padova (Italy) on April 1st, 1964. He received the Master degree (cum laude) in Electronic Engineering and the Ph. D. degree in Electronics and Telecommunications Engineering from the University of Padova in 1989 and 1993, respectively. From 1993 to 1994, he was lecturer with the Department of Mathematics and Statistics, University of New Mexico, Albuquerque. From 1994 to 1998, he was Assistant Professor with the Department of Electronics and Informatics, University of Padova. In 1997 he has been appointed as Visiting Researcher at the University of Limoges. Since 1998 he is with the University of Brescia, where he is Full Professor of Electromagnetic Fields. In 2010 and 2011 he has been appointed as Visiting Professor at the Massachusetts Institute of Technology. His technical interests are in terahertz technologies, graphene photonics, nanophotonics and optical antennas, in the linear and non-linear regimes. He has authored or coauthored more than 300 among papers and conference contributions and he is and has been principal investigator in several European and national funded research projects.



VALERIO F. GILI went to the University of studies of Rome “La Sapienza”, where he studied Physics and took his bachelor degree in 2012. He received the Master degree in Physics at the University of studies of Rome in 2014, doing an experimental thesis on quantum optics with Paolo Mataloni and Fabio Sciarrino. He now attends PhD. Study in the Laboratoire Matériaux et Phénomènes Quantiques at Paris Diderot University.



GIUSEPPE LEO (born in Italy in 1966, married, two children) received his MS summa cum laude in EE at La Sapienza University (Rome, 1990) and his PhD in Physics at Paris-Sud University (2001). Since 1992, he was assistant professor at Roma-Tre University, where he became associate professor in 2002. Since 2004 he has been full professor at Paris Diderot University. G. Leo’s research is in optoelectronics and nonlinear optics, with a focus on AlGaAs integrated optics. Since 2005, he has directed 4 post-docs and 6 PhD students. Previously, he had supervised 12 Laurea theses in Italy. He has published 67 articles on peer-reviewed journals, 9 book chapters and about 110 conference papers. Finally, he has edited 1 book and registered 2 patents. Head of the DON group of the MPQ Laboratory, in 2006 he has founded and directed 2 Professional Schools at Paris Diderot (“Materials analysis” and “Biophotonics”). Since 2008 he has led the foundation of the Denis Diderot School of Engineering, of which he is presently the director. Over the last years, he has managed 5 bilateral programs and participated in 3 RTD projects and 1 RTN network. He is presently the coordinator of the FP7 FET TREASURE Project.

ABSTRACTION ON NETWORK MODEL UNDER INTEROPERABLE SIMULATION ENVIRONMENT

Bong Gu Kang, Byeong Soo Kim, and Tag Gon Kim
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, 305-701, Republic of Korea
E-mail: kbgmode@kaist.ac.kr

KEYWORDS

Abstraction of network model, interoperable simulation environment, metamodeling, simulation execution time.

ABSTRACT

This paper proposes a method for abstraction of a network model that enables a reduction of simulation execution time under the interoperable military system simulation environment that consists of two models: the network model and computer-generated forces (CGF) model. This paper illustrates the 1) overall procedure of abstraction and 2) empirical analysis. In the abstraction step, our approach uses the information between the two models and, as well as of the network model, unlike previous research. To show how the procedure can be applied, we first implement the CGF model, network model, and interoperable simulation environment. Then, we apply our method to the simulation environment in a case study by comparing with previous study in terms of accuracy and speed. From the empirical results, we can draw the conclusion that the accuracy of the simulation is significantly enhanced by sacrificing a little execution time within an acceptable range. In closing, we expect that this approach will help people conduct replication simulations that require long execution times in one execution.

INTRODUCTION

In modern warfare, the influence of communication is considered as important as other factors, which is why the word “communication (C)” is considered as one element of Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR). In this respect, many researchers in defense modeling and simulation (M&S) have paid close attention to the effects of communication in defense-system simulations. Nevertheless, until recently, many military models have assumed sufficient communication functions between entities in the military model. To be specific, many models still assumed perfect communication, that is, no delay and loss of communication (Kim et al. 2012; Sung and Kim 2012); others assumed a simple connection model with probability (Yang et al. 2006) or a relatively more complex model (Shin et al. 2013) by designing the model as one sub-model of the entire military model to determine whether messages were transmitted. By extension, others tried to depict the effects of communication by doing

integration or interoperation with existing standalone communication, network models developed through various tools suitable for communication modeling and the military model, or the CGF model. (Walsh et al. 2005; Paz and Baer 2008; Kang and Kim 2013).

Although the integration or interoperation approach has an advantage to be able to depict the detailed effects of communication, each requires long execution times to complete the simulation, owing to the complexity of the network model. Thus, these approaches make it difficult to experiment on a large number of scenarios and replications. To overcome this problem, Porche et al. (2004) constructed an abstracted model of the network model and integrated the abstracted model with the CGF model, or war-game model, in order to measure the impact of communication effects on combat power (Porche et al. 2006). In this procedure, the abstracted model unfortunately considered only inputs of the network model, such as specification of communication equipment, not information between the network model and CGF model, thereby reducing the accuracy of the abstracted model.

To face the above challenge, this study presents a method that considers the information between two models: the network model and CGF model, as well as the inputs of the network model. This paper illustrates 1) the overall procedure composed of three phase—data acquisition from simulation, abstracted model construction from the data, and application of the abstracted model in analysis—and 2) empirical simulation results in order to show how the procedure can be applied in defense simulation. In detail, we first construct an infantry-company-level CGF model and a network model including a depiction of the mobile ad-hoc network (MANET) using ad hoc on-demand distance vector routing (AODV) protocols in the battlefield (Kaur and Sharma 2013). Then, we implement an interoperation simulation environment using the above two models based on high-level architecture (HLA), a specification for interoperation of distributed heterogeneous models (IEEE Standards Association 2010). Finally, the network model under the interoperation environment is abstracted according to our proposed procedure, and it is compared with previous studies (Porche et al. 2004, 2006) with regard to the accuracy and simulation execution speed of the model.

This study is organized as follows: In the next section, related works on abstraction of the network model are described. Then, we illustrate our method to overcome

related work's flaws and its application. Finally, we evaluate our proposed work and draw a conclusion.

RELATED WORKS

Because M&S has been regarded as an important technique in the communication research fields, various open-source, commercial communication, or network models focusing on a network analysis have been developed (Pan and Jain 2008). Such models usually calculate packet-delivery ratio (PDR) and end-to-end delay as their outputs from simulation. The former means the ratio of the number of delivered packets from the source to the destination node, and the latter indicates the average time taken by packets to arrive at the destination from the source node. For this reason, earlier studies considered the two factors (i.e., PDR and end-to-end delay) as outputs of an abstracted network model (Porche et al. 2004).

As mentioned before, the target-simulation model focusing on communication effects in the battlefield consists of two models, as shown in the left part of Figure 1: the CGF and network model. In detail, the combat entities of the CGF model and nodes (i.e., communication equipment) of the network model share their positions to represent them as deployed under the same battlefield condition. To be more specific, when communicating between two entities in the CGF model, the CGF model first sends a packet to the network model. After that, the communication effects are calculated based on the PDR and end-to-end delay in the network model, and the packet is again delivered to the CGF model. Eventually, packet loss and delay affect the combat power of the CGF model, and thus, the influence of the network parameter, such as the transmission power of the communication equipment, on combat power can be measured and analyzed.

Under this circumstance, the network model requires long execution times due to high complexity. Thus, it causes the necessity of abstraction of the model. The right part of Figure 1 shows how the network model can be abstracted. An earlier study (Porche et al. 2004) only considered the network parameter as its inputs by not considering network information such as node positions. However, this approach has a limitation in showing high accuracy.

To be specific, the communication between nodes over long distances is conducted by multi-hops using intermediate nodes in the real world; in spite of that, the earlier study cannot express such a situation. Thus, this paper proposes a method that is able to consider not only the network information from the CGF model but also the network parameters of the network model for enhanced accuracy.

PROBLEM DEFINITION

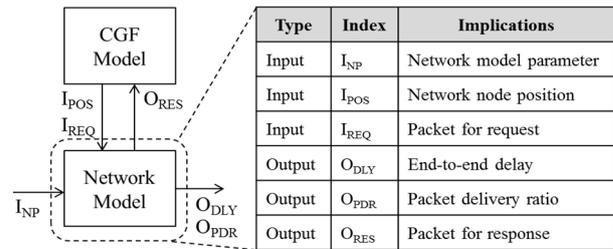


Figure 2: Input and Output of Network Model

Figure 2 illustrates more specific inputs and outputs of the target network model for abstraction, which is mentioned in the left part of Figure 1. Input consists of 3 factors: network model parameter, node position, and packet for request. Output also has 3 factors: PDR, end-to-end delay, and packet for response. The packet for response is calculated with the PDR and end-to-end delay of the network model against arrived packets for request. The PDR and end-to-end delay can vary due to the network model parameter and node position. In terms of such a network model, this paper focuses on constructing an abstracted network model, including the same input/output factors, assuming that we can only access input/output, not inner state. When it comes to the packet for response, which is one of the outputs, it is indirectly acquired by the PDR and end-to-end delay against the arrived request packet, not directly calculated. In other words, the packet for response can be calculated only if we know the PDR and end-to-end delay. Thus, our method only focuses on the abovementioned two outputs. Even though the proposed abstracted method explains the two

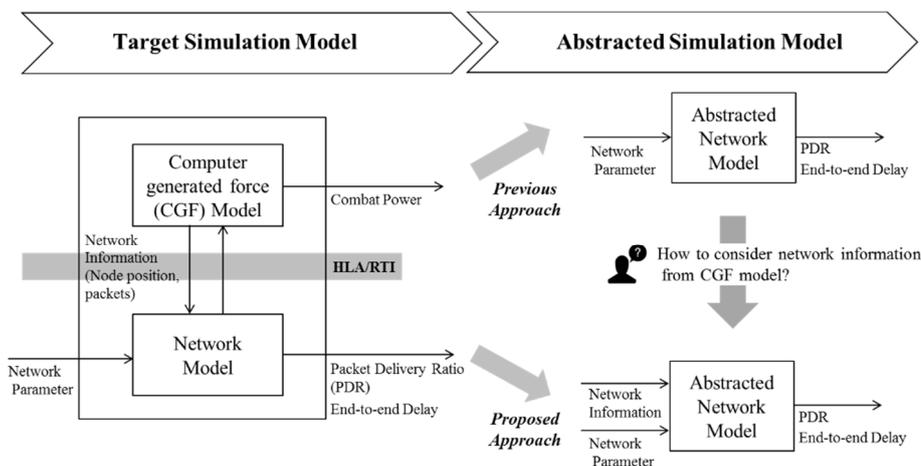


Figure 1: Difference between Previous and Proposed Approach on Abstraction of Network Model

outputs, our abstracted model can be substituted for the existing network model because our model internally includes a function that reflects the PDR and end-to-end delay on the packet for request and transmits the packet for response to the CGF model by implementing the function in the abstracted network model.

PROPOSED WORKS

This section presents an overall procedure to show how the network model can be abstracted and how it can be used with the CGF model for interoperable simulation. Before describing the procedure, we describe prior knowledge to aid understanding the process.

In the previous section, we confirmed that node position is one of the outputs, and it has effects on communication. Thus, the positions of all nodes should be reflected to enhance the accuracy, not only one node position, which receives from the CGF model because the positions of all nodes include other nodes' positions besides a source and destination node. However, to represent all node positions in this way expands the dimension of inputs when the number of nodes is increased, and it needs much more time to execute simulations based on the number of nodes. Also, these absolute positions of nodes cannot give any other insight. To confront this weakness, we condensed the position of all nodes to distance and network density, as shown in Figure 3, which have been considered major factors in the communication research field (Adam et al. 2010). The former represents the distance between the source and destination node, and the latter represents the number of nodes in the rectangular area by the source and destination node. To put it concretely, a long distance requires more hops for communication, and it consequently causes the increased end-to-end delay and the decreased PDR; on the other hand, a larger network density implies the probability of the existence of intermediate nodes between the source and destination node, which causes the increased end-to-end delay and PDR. Due to this importance of these facts, we chose the two factors: distance and network density. Whenever I_{POS} from the CGF model occurs, one must save the position and then calculate the two factors whenever I_{REQ} is sent from the CGF model.

Figure 4 illustrates the overall procedure of the network model abstraction, and the process consists of 3 phases: 1) simulation using the target network model, 2) abstraction, (i.e., metamodeling) of the network model, and 3) simulation and analysis on an abstracted network model.

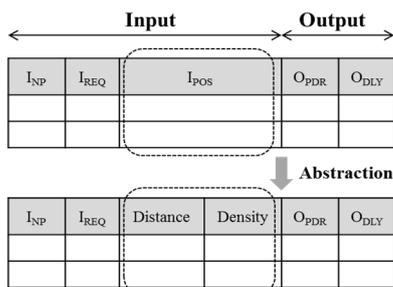


Figure 3: Identify Input for Abstracted Network Model

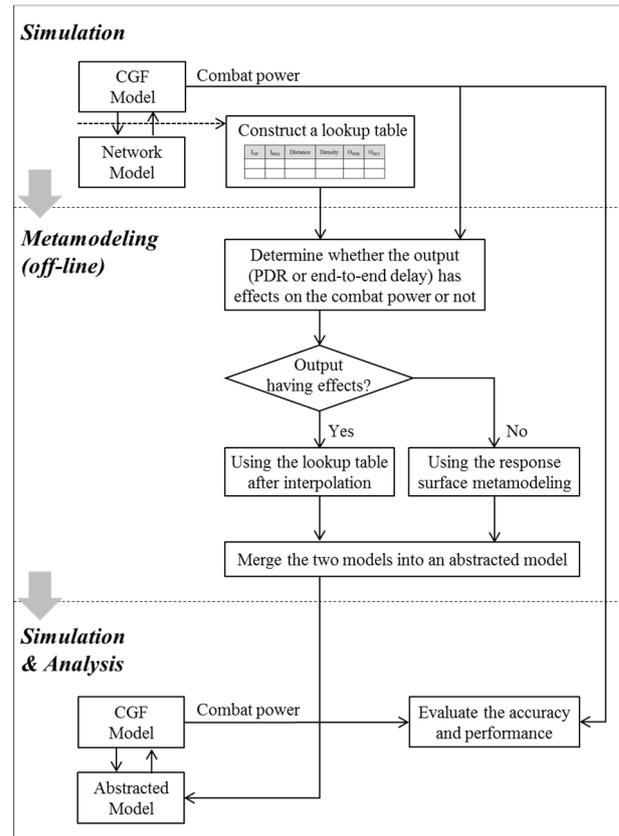


Figure 4: Overall Process of Abstracted Network Model

In the first step, interoperable simulation using the CGF and target network model is executed. During this time, data related to communication is saved, as in the form of Figure 3. Also, combat power is measured as an output of the CGF model.

After finishing the simulation, the abstraction process is conducted with data acquired in the first stage. In this procedure, we first determine whether the output of the network model (O_{PDR} , O_{DLY}) has effects on the combat power. If the output has no impact on the combat power, we don't need to represent the output in detail. Figures 6 and 7 show the decision process, which will be mentioned in the case-study section. In other words, if the output of the network model has an effect on the result of the CGF model, the output should be specifically depicted. Thus, this study uses the tabulation technique including interpolation, which was recently used in defense M&S to abstract one system between two (Bae et al. 2016). On the contrary, if the output of the network model has little effect on the result of the CGF model, it does not need to be described in detail. For this reason, this paper represents it as a simple model using a general regression method, such as response surface metamodeling (RSM; refer to Table 2 in the case-study section). After constructing each model against the output (O_{PDR} , O_{DLY}), we merge the two models into an abstracted model, since the target network model is one model component. In the last phase, the target network model is substituted for the abstracted model and participates in simulation with the CGF model. From this simulation, the abstracted

network model can be evaluated in terms of accuracy and speed. The fundamental objective of using the target network model is to derive the combat power of the CGF model, which reflects effects of communication, not to merely measure its functions or performance (although it goes without saying that it is ideal to depict the functions perfectly). In this regard, the accuracy should be evaluated against the output of the CGF, not the network model. Thus, the combat power of experimental results using the target network model and abstracted network model should be measured and compared. Regarding simulation speed, this paper measures the simulation execution, or run time, since the ultimate goal of using the abstracted model is to shrink the execution time. After acquiring the two performances: accuracy and speed, this paper analyzes the trade-off between the two performances by comparing with the previous method and determining whether the proposed method is valid.

EXPERIMENT

This section illustrates a case study for the proposed method in the following order: simulation model design, experiment design to compare the simulation model and its abstraction, and experiment results.

Simulation Model Design

The simulation model for abstraction consists of two models, the CGF and network model, and conducts the army's ground operations at infantry-company level. Recently, owing to the importance of communication, the equipment for communication is allocated to each soldier, which is reflected in the defense M&S (Kuosmanen 2002) domain. In this respect, we first constructed 131 combat entities—108 soldiers, 12 squad command and control (C2)s, 4 platoon C2s, 1 company C2, and 6 mortars—for the CGF model using discrete event systems specification (DEVS) formalism, which has been widely used for defense M&S (Zeigler et al. 2000; Seo et al. 2011). Then, we also constructed 131 network equipment, which corresponded with the above combat entities of the CGF model in the network model. We assumed that the network equipment used AODV protocol for MANET, which has been described as the form of a finite state machine (FSM) (Wagner et al. 2006). Finally, to depict the combat entities and network equipment as under the

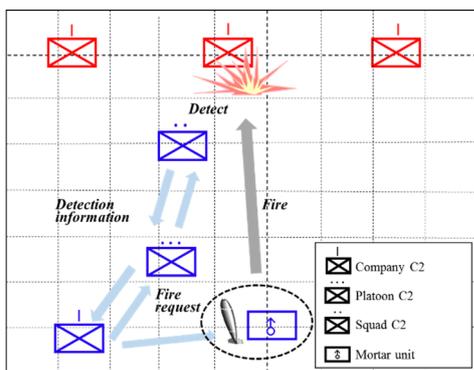


Figure 5: Illustration of Abstracted Combat Scenario

same battlefield situation, we participate the two models in an HLA-based interoperable simulation through runtime infrastructure (RTI) by using one object and one interaction for each node's position information and packet information between two nodes, respectively.

Figure 5 depicts a combat scenario that simplifies the complex hierarchy of the CGF model. Concretely, the blue force's combat entities and equipment were located in a $2 \text{ km} \times 2 \text{ km}$ operation area, and one company within the blue force executed defense operations against three companies within the red force. When the red force approached from the upper region of the operation area, a blue soldier detected the red one and hierarchically sent the detection information to the top company C2 via squad and platoon C2 in order (Dekker 2002). After that, the company C2 conducted a threat evaluation and weapons, such as mortar allocation, and then transmitted a fire request message to the mortar and soldier through a subordinate C2, as shown by the light blue arrows in Figure 5. In such a transmission of the messages, the communication effects are reflected and affect the combat power. For instance, an in-direct attack using a mortar will be smooth if the condition of communication is perfect (i.e., no packet delay and loss), and it can consecutively enable an easier direct attack by soldiers.

Experiment Design

The objective of this experiment is to compare existing and proposed methods for abstraction against two factors: accuracy and simulation speed. This paper chose enemy survivability rate and simulation execution, or run time, as performance indexes against the accuracy and speed, respectively. Also, we considered net diameter in MANET AODV as a parameter of the network model, as shown in Table 1, which shows the maximum possible number of hops between two nodes in the network. Against this parameter, we chose 10 experimental points and ran repeat simulations 30 times per one experiment point.

Finally, the simulation environment for this case study is as follows. For the CGF model, CPU: I5-3550 3.3 GHz, RAM: 4 GB, DEVSim++ v.3.1, and KHLAAdaptor were used (Kim et al. 2011). In the network model, OPNET v.14.5 was used under the same hardware. These models, or simulators, were interoperated by RTI 1.3-NG. The simulation progressed over 2 hours (i.e., when the enemy survivability was sufficiently saturated).

Table 1: Network Model Parameter

Parameter name	Parameter description	Parameter level
Net Diameter	Maximum possible number of hops between two nodes in the network	1, 2, ..., 10 (10 cases)

Experiment Result

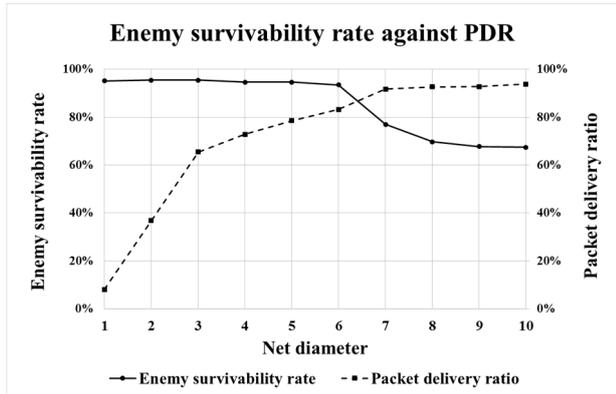


Figure 6: Enemy Survivability Rate against PDR

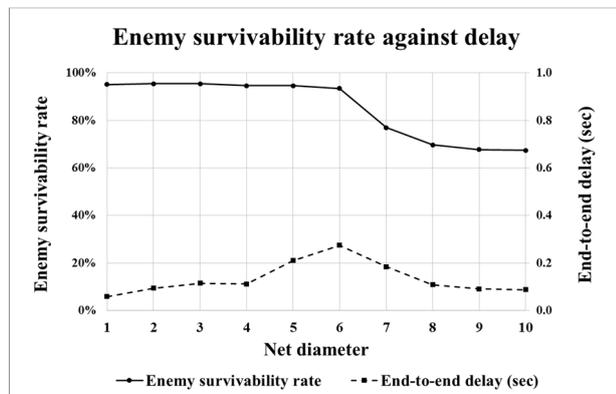


Figure 7: Enemy Survivability Rate according to End-to-end Delay

Figures 6 and 7 show experimental results when using the network model without abstraction. We can identify that the enemy survivability rate has a declining tendency when the net diameter increases from the above two figures. The reason for this is that the increased net diameter enables multi-hop communication. In this circumstance, and by extension, we are able to profoundly analyze the effects of communication performance on the enemy survivability rate by observing the PDR and end-to-end delay in each figure. For example, Figure 6 shows that the PDR has numerous effects on the enemy survivability rate; on the other hand, we can recognize from Figure 7 that the end-to-end delay has little effect on the survivability rate because the delay is too small.

Table 2: Abstraction Model through Previous Method

Abstraction model input/output	Abstraction model identified parameters
PDR =f(x=Net Diameter)	$0.260690x - 0.016206x^2 - 0.094425$, $R_{adj}^2 = 0.955$
End-to-end delay =g(x=Net Diameter)	$0.077975x - 0.006822x^2 - 0.033217$, $R_{adj}^2 = 0.509$

On the basis of the data acquired by the network model, we constructed two kinds of abstraction models using a previous (Porche et al. 2004) and a proposed method.

For the former's case, we constructed an abstraction model covering two functions, PDR and end-to-end delay, against the network model parameter using the RSM method, as shown in Table 2.

In the case of the latter, we constructed another abstraction model representing the PDR and end-to-end delay with tabulation technique and RSM, respectively. Since we recognize that the end-to-end delay has little effect on the enemy survivability rate from above the experiment, we chose the RSM because it was relatively simple compared to the tabulation technique. Meanwhile, we made a table consisting of 3 inputs—net diameter, distance, and density, against the PDR—and then we set quantization size as 1, 20, and 1 in each input and used linear interpolation for tabulation.

After constructing two abstracted models as mentioned above, we substituted the network model to the abstracted model and executed the simulation. Figure 8 shows simulation results in terms of accuracy. From Figure 8, we observed that the enemy survivability rate from the previous method deviated far more than the proposed method. For the quantitative analysis, we measured the root mean square error (RMSE) of the survivability rate. The RMSE of the case using the existing method-based model and the network model was measured to be 14.16%; on the other hand, the RMSE was recorded as 4.58% when using the proposed method-based model and the network model. This indicates that the proposed method was 3.09 times more accurate than the earlier study. Regarding speed, we measured a total execution time of about 300 trials (10 experimental points \times 30 replications), and the time was recorded as 111,658, 459, and 493 minutes against cases using the network model, previous model, and proposed abstracted network model, respectively. This result shows that 1.07 times more execution time is required when using the proposed method compared with the existing one.

To sum up the experimental results, accuracy is considerably enhanced with the proposed method without exhausting the corresponding execution time.

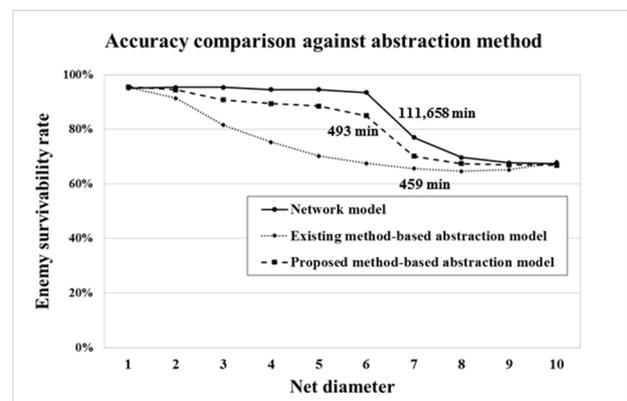


Figure 8: Accuracy Comparison according to the Abstraction Methods

CONCLUSIONS

In modern warfare, the importance of communication has increased. Thus, many M&S researches have used complex network models, such as simulation tools that are dedicated to the network domain, to analyze the influence of communication effects on combat power in the military system. However, to use the complex network models unfortunately causes long execution time problems due to the complexity of the network models.

To overcome this weakness, many studies have tried to use the network model in an abstracted form, and use it with the computer generated forces (CGF) model by integrating/interoperating the two models. Nevertheless, the previous research did not consider the information between the two models during the abstraction process, and it consequently caused an accuracy decrease.

In this study, we presented a method that considers the information between two models, as well as the inputs of the network model, like property of network equipment. For this, we described the overall procedure of the abstraction using various abstraction techniques and showed empirical simulation results acquired through a case study by comparing the proposed method with the previous study.

As a main contribution, the accuracy of the abstracted model was significantly enhanced by sacrificing a little simulation execution time within an acceptable range. Furthermore, this proposed method can help people conduct simulation-based analyses against various scenarios and replications within time constraints.

ACKNOWLEDGEMENT

This work was supported by the Defense Acquisition Program Administration and Agency for Defense Development under the contract UD140022PD, Korea.

REFERENCES

- Adam N., M.Y. Ismail, and J. Abdullah. 2010. "Effect of node density on performances of three MANET routing protocols." In Proceedings of the 2010 Electronic Devices, Systems and Applications Conference (Kuala Lumpur, Malaysia, Apr.11-14), 321-325.
- Bae J.W., J.H. Kim, I.C. Moon and T.G. Kim. 2016. "Accelerated simulation of hierarchical military operations with tabulation technique." *Journal of Simulation* 10, No.1 (Feb), 36-49.
- Dekker A.H. 2002. C4ISR architectures, social network analysis and the FINC methodology: an experiment in military organisational structure. Information Technology Division Electronics and Surveillance Research Laboratory, Australia. (Jan)
- IEEE Standards Association. 2010. 1516-2010 IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) – Framework and Rules.
- Kang B.G. and T.G. Kim. 2013. "Reconfigurable C3 simulation framework: interoperation between C2 and communication simulators." In Proceedings of the 2013 Winter Simulation Conference (Washington D.C., Dec.8-11). IEEE, Picataway, N.J., 2819-2830.
- Kaur S. and C. Sharma. 2013. "An Overview of Mobile Ad hoc Network: Application, Challenges and Comparison of Routing Protocols." *IOSR Journal of Computer Engineering (IOSR-JCE)*, e-ISSN, 2278-0661.
- Kim J.H., I.C. Moon, and T.G. Kim. 2012. "New insight into doctrine via simulation interoperation of heterogeneous levels of models in battle experimentation." *Simulation* 88, No.6 (Jun), 649-667.
- Kim T.G., C.H. Sung, S.Y. Hong, J.H. Hong, C.B. Choi, J.H. Kim, K.M. Seo and J.W. Bae. 2011. "DEVSim++ toolset for defense modeling and simulation and interoperation." *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 8, No.3 (Nov), 129-142.
- Kuosmanen P. 2002. Choosing routing protocol for military ad hoc networks based on network structure and dynamics. Master's Thesis, Helsinki University of Technology.
- Pan J. and R. Jain. 2008. A survey of network simulation tools: Current status and future developments. Washington University, St. Louis, Mo.
- Paz B.D. and J.A. Baer. 2008. "Communication effect server integration with OneSAF for mission level simulation." In Proceedings of the 2008 Fall Simulation Interoperability Workshop (Orlando, FL, Jul.12-14).
- Porche I, L. Jamison, and T. Herbert. 2004. Framework for Measuring the Impact of C4ISR Technologies and Concepts on Warfighter Effectiveness Using High Resolution Simulation. Rand, Santa Monica, Ca. (Jun)
- Porche III, R. Isaac, and W. Bradley. 2006. The impact of network performance on warfighter effectiveness. Rand, Santa Monica, Ca.
- Seo K.M., H.S. Song, S.J. Kwon and T.G. Kim. 2011. "Measurement of effectiveness for an anti-torpedo combat system using a discrete event systems specification-based underwater warfare simulator." *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 8, No.3, (Nov), 157-171.
- Shin K.H., H.C. Nam, and T.S Lee. 2013. "Communication modeling for a combat simulation in a network centric warfare environment." In Proceedings of the 2013 Winter Simulation Conference (Washington D.C., Dec.8-11). IEEE, Picataway, N.J., 1503-1514.
- Sung C.H. and T.G. Kim. 2012. "Collaborative modeling process for development of domain-specific discrete event simulation systems." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 42, No.4 (Jul), 532-546.
- Wagner, F.; R. Schmuki; T. Wagner; and P. Wolstenholme. 2006. Modeling software with finite state machines: a practical approach. CRC Press, Boca Raton, F.L.
- Walsh J., J. Roberts, and W. Thompson. 2005. "NCW End-to-end (NETE) model for future C2 architecture assessments." In Proceedings of the 2005 International Command and Control Research and Technology Symposium (McLean, VA, Jun.13-16).
- Yang A., H.A. Abbass, and R. Sarker. 2006. "Land combat scenario planning: A multi objective approach." In *Simulated Evolution and Learning 2006*, T.D. Wang, X. Li, S.H. Chen, X. Wang, H. Abbass, H. Iba, G.L. Chen, and X. Yao (Eds.). China, Hefei, 837-844.
- Zeigler, B.P; H. Praehofer; and T.G. Kim. 2000. Theory of modeling and simulation. Academic Press, Orlando, F.L.

AUTHOR BIOGRAPHIES

BONG GU KANG is a PhD. 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), defense modeling and simulation, and interoperation simulation.

BYEONG SOO KIM is a PhD. 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.

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 has published about 200 papers on M&S theory and practice in international journals and conference proceedings. He is very active in defense modeling and simulation in Korea.

MODELS AND ALGORITHMS FOR ABILITIES EVALUATION OF ACTIVE MOVING OBJECTS CONTROL SYSTEM

Boris Sokolov
¹St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14th line 39, St.Petersburg, 199178, Russia
²ITMO University
49 Kronversky Pr., St.Petersburg, 197101, Russia
sokolov_boris@inbox.ru

Vladimir Kalinin
Military-space academy named after A.F.Mozhaisky
13 Gdanovskaya str., St. Petersburg, 197198, Russia
kvn112@mail.ru

Sergey Nemykin
The Design Bureau "Arsenal" named after M.V.Frunze
1-3 Komsomola street, 195009, St. Petersburg, Russia
kbarsenal@kbarsenail.ru
Complete Address Line 4

Dmitry Ivanov
Berlin School of Economics and Law
Department of Business Administration
10825, Berlin, Germany
Dmitri.ivanov@mail.ru

KEYWORDS

Active moving objects control system, evaluation of goal abilities, attainability set and scheduling, optimal control program.

ABSTRACT

One of the important problems in active moving objects control system (AMO CS) is the evaluation of goal abilities, i.e., potential of the system to perform its missions in different situations. Thus, the preliminary analysis of information and technological and goal abilities (GA and ITA) of AMO CS is very important in practice and can be used to obtain reasonable means of the AMO exploitation under different conditions. In the paper models and algorithms for abilities evaluation of AMO CS are proposed.

ABBREVIATIONS AND NOTATION

AMO — active moving objects
AS — attainable sets
CS — control system
CTS — complex technical systems
GA — goal abilities
ITA — informational and technological abilities
IZ — interaction zone
NFDDS — nonstationary finite-dimensional differential dynamic systems
OS — object-in service
OPS — optimal program control
SDC — structure-dynamics control
 $D(t_f, t_0, \mathbf{x}(t_0))$ — attainable sets (AS)

$D^-(t_f, t_0, \mathbf{x}(t_0))$ — approximation of AS

$\mathbf{x}(t_0)$ — initial state vector of AMO CS

$\mathbf{x}(t_f)$ — end state vector of AMO CS

t_0 — initial point of time

t_f — final time of the scheduling interval

$\mathbf{u}(t)$ — control vector, represents AMO control program

$\varepsilon_{ij}(t)$ — element of the present time function of time-spatial constraints ($\varepsilon_{ij}(t) \in \{0, 1\}$)

INTRODUCTION

The general objects of our investigation are active moving objects control system (AMO CS). The notion "Active Mobile Object" generalizes features of mobile elements dealing with different complex technical systems (CTS) types (Kalilin et al., 1985, Okhtilev et al., 2006, Ivanov et al., 2010). Depending on the type of CTS Active Mobile Objects can move and interact in space, in air, on the ground, in water, or on water surface. Active Mobile Object can be regarded as multi-agent system. We distinguish two classes of AMO. AMO-one, namely AMO of the first type. This type of AMO fulfills CTS principal tasks. AMO-two supports functioning of AMO-one. Objects-in-service (OS) can be regarded as external AMO. Analysis of the main trends of modern AMO CS indicates their peculiarities such as: multiple aspects and uncertainty of behavior, hierarchy, structure similarity and surplus for main elements and subsystems of AMO CS, interrelations,

variety of control functions relevant to each AMO CS level, territory distribution of AMO CS components.

One of the main features of modern AMO CS is the variability of their parameters and structures as caused by objective and subjective factors at different phases of the AMO CS life cycle (Klir 2005, Okhtilev et al., 2006, Ivanov et al., 2010). In other words, we always come across the AMO CS structure dynamics in practice. Under the existing conditions the AMO CS potentialities increment (stabilization) or degradation reducing makes it necessary to perform the AMO CS structures control (including the control of structures reconfiguration). There are many possible variants of AMO CS structure dynamics control. For example, *they are alteration of AMO CS functioning means and objectives; alteration of the order of observation tasks and control tasks solving; redistribution of functions, of problems and of control algorithms between AMO CS levels; reserve resources control; control of motion of AMO CS elements and subsystems; reconfiguration of AMO CS different structures.*

According to the contents of the structure-dynamics control problems belong under the class of the AMO CS structure-functional synthesis problems and the problems of program construction, providing for the AMO CS development.

One of the important problems in AMO CS structure-dynamic control is the evaluation of goal abilities, i.e., potential of the system to perform its missions in different situations. Thus, the preliminary analysis of information and technological and goal abilities (GA and ITA) of AMO CS can be used to obtain reasonable means of the objects B_j , $j = 1, \dots, m$ exploitation under different conditions. The numerical estimations of AMO CS GA and ITA should be based on the system of measures. These measures can be regarded as characteristics of AMO CS potential effectiveness. The GA measures characterizing different levels of AMO CS are interrelated and have a hierarchical structure. The leading role of information and technological aspects of the goal-abilities (GA) evaluation is a result of the influence of the technology structure (the structure of AMO CS control technology) upon the other AMO CS structures (organizational structure, technical structure, etc.) So the information and technological abilities (ITA) of a system ought to be evaluated first of all. These abilities can be measured as AMO CS capacities. The following measures are to be evaluated: the total number of objects in a given macro-state over a fixed time period or at a fixed point of time; the total number of working operations performed over a fixed time period σ or by the time point t .

Parallel with the enumerated measures of ITA the following measures of GA can be used the total possible number of objects-in-service (OS) over the time period σ ; the total time that is necessary for the execution of all interaction operations with OS. If the uncertainty factors are considered (the stochastic, probabilistic, or fuzzy models can be applied) the measures of GA can be

evaluated as the expectation (or the fuzzy expectation) of the number of serviced objects by a given time point; the probability (its statistical estimation) of successful service for the given objects. Similar measures can be proposed for ITA estimations, for example the expectation of the number of objects in a given macro-state, the probability of technological operations fulfillment.

The problem of AMO CS GA and ITA evaluation and analysis can be solved on the basis of structure dynamics control models (the model M and its components Mo , Mk , Mn , Me , Mg , Mv , Mc , Mp) (Okhtilev et al., 2006, Kalinin et al., 1985, 1987).

RESEARCH METHODOLOGY

The proposed approach is based on fundamental scientific results of optimal program control (OPC) theory regarding dynamic interpretation of scheduling problems and performance evaluation. The research methodology is based on the following basic principles.

The first feature of the research methodology is an original dynamic representation of AMO CS schedule as OPC control vector $\mathbf{u}(t)$, represents AMO control programs (plans of AMO CS functioning) (Kalinin et al., 1985, 1987). The AMO CS scheduling is interpreted as dynamic process of operations control. From these points of view, the understanding of "dynamic scheduling" in this control theoretic study differs from the concept of dynamic scheduling in traditional rescheduling techniques (compare with Vieira et al. 2000). The advantages of the scheduling with the help of OPC have been extensively discussed in (Khmelnitsky et al., 1997, Ivanov and Sokolov 2010).

For the stage of AMO CS scheduling, we formulate the OPC model as a linear non-stationary finite-dimensional controlled differential system with the convex area of admissible control. Such a model form is favourable because of (i) possibilities to calculate OPC and (ii) to approximate attainable sets (see further in this paper).

The calculation procedure for OPC is based on the Pontryagin's maximum principle. With representing the AMO CS schedule as OPC, it becomes possible to perturb the parameters of AMO CS schedule at any point of time and of different severity (e.g., in the interval from zero to full resource breakdown) and to reflect non-stationary perturbations in the further calculation of robustness metric. Hence, the parameters and their variations in dynamics are explicitly expressed in the scheduling model and can be used for the robustness analysis in order to integrate the robustness objective as a non-stationary performance indicator in AMO CS scheduling.

The second feature of the research methodology is the dynamic representation of AMO CS schedule execution under different uncertainties based on attainable sets (AS). An AS of a controllable dynamic system in the state space is typically notated as $D(t_f, t_0, \mathbf{x}(t_0))$, where

t_0 , is an initial point of time, $\mathbf{x}(t_0)$ is an initial state vector, and t_f is the final time of the interval.

The AS at current time $t_1 \in (t_0, t_f]$ includes all points of the system's state trajectories (e.g., a set of all possible execution scenarios which may occur for the AMO CS schedule after the perturbations) at time t_1 under the following conditions: each trajectory begins at time $t = t_0$ in the state $\mathbf{x}(t_0)$ and is formed through some allowable variations of control $\mathbf{u}(t_0)$ within the time interval $(t_0, t_f]$ (Gubarev et al. 1988, Chernousko 1994, Clarke et al. 1995, Okhtilev et al. 2006). It is important that the AS concept be applicable to multi-step procedures. It may be possible to derive the multiple decoupled attainable sets at each point of time that ensure that the overall schedule meets the performance requirements as long as the constituent steps are operated within the AS.

In less technical words, the AS approach is to determine a range of operating policies (the union of which is called as an AS) during the scheduling stage over which the system current performance can be guaranteed to meet certain targets, i.e., the output performance. Basically the AS is a fundamental characteristic of any dynamic system. If the AS is known its basic characteristics in essence replace with themselves all the information necessary about system dynamics, the stability of its functioning and output performance. The AS characterizes all possible states of the AMO CS schedule subject to different variations of AMO CS parameters in nodes and channels (e.g., resource capacity availability).

Besides, if the AS is known, it becomes possible to analyse the dependence between the scheduling results subject to output performance (e.g., service level and delivery reliability) and the structure and properties (e.g., inventory quantity and location, lot-sizes, transportation channels and the intensity of their usage) of the start and end states $(X_0, X_f]$. In other words, it becomes possible to define the area in which permissible solutions (e.g., AMO CS schedules) are included. On the other hand, the AS analysis may show that, with the given resource and at the given time horizon, it is impossible to achieve the required output performance; hence, we should introduce additional resources or expand the supply cycle.

MODELS AND ALGORITHMS FOR ABILITIES EVALUATION OF ACTIVE MOVING OBJECTS CONTROL SYSTEM

The problem of AMO CS GA and ITA evaluation and analysis can be solved on the basis of structure-dynamics control models (the model M and its components $Mo, Mk, Mn, Me, Mg, Mv, Mc, Mp$). (Okhtilev et al.,2006, Ivanov and Sokolov 2010).

These models have a form of nonstationary finite-dimensional differential dynamic systems (NFDDS)

with reconfigurable structures. So the problem of GA and ITA evaluation can be regarded as a problem of NFDDS controllability analysis. The latter problem, in its turn, can be solved by the NFDDS attainability set $D(t_f, t_0, \mathbf{x}(t_0))$ construction. If the attainability set (AS) is obtained, the solvability of the previously stated boundary problems for structure-dynamics control (SDC) can be checked in accordance with the sets of initial X_0 and final X_f states ($\mathbf{x}(t_0) \in X_0, \mathbf{x}(t_f) \in X_f$), with the considered period of time, with time-spatial, technical, and technological constraints. Moreover, the problems of AMO CS GA and ITA evaluation and analysis can be formulated as follows:

$$J'_{06}(\mathbf{x}(\cdot)) \rightarrow \min_{\mathbf{x}(\cdot) \in D(t_f, t_0, \mathbf{x}(t_0))}, \quad (1)$$

where $D(t_f, t_0, \mathbf{x}(t_0))$ is the attainability set of the dynamic system (model) M ; $J'_{06}(\mathbf{x}(\cdot))$ – is the initial functional transformed to the form of Mayer's functional. It is important that the alteration of objective functional does not imply the recalculation of the attainability set $D(t_f, t_0, \mathbf{x}(t_0))$. If the dimensionality of AMO CS GA and ITA evaluation and analysis problems is high, then the construction of the attainability sets becomes a rather complicated problem. Therefore, the approximations of $D(t_f, t_0, \mathbf{x}(t_0))$ ought to be used (Chernousko F.L. 1994). Now we introduce the algorithm of $D(t_f, t_0, \mathbf{x}(t_0))$ construction. The boundary points of the set $D(t_f, t_0, \mathbf{x}(t_0))$ are obtained as the solutions of the optimal control problems (Chernousko F.L. 1994, Okhtilev et al.,2006, Ivanov et al., 2010):

$$J''_{06}(\mathbf{x}(\cdot)) = \mathbf{c}^T \mathbf{x}(t_f) \rightarrow \min_{\mathbf{u} \in Q_p(\mathbf{x})}, \quad (2)$$

where \mathbf{c} is a vector such that $|\mathbf{c}| = 1$. For a given vector \mathbf{c} we obtain the optimal control $\mathbf{u}^*(t)$, the appropriate state vector $\mathbf{x}^*(t_f)$ that is equal to some boundary point of $D(t_f, t_0, \mathbf{x}(t_0))$, and the hyperplane $\mathbf{c}^T \mathbf{x}^*(t_f)$ to $D(t_f, t_0, \mathbf{x}(t_0))$ at the point $\mathbf{x}^*(t_f)$.

Let $\bar{\Delta}$ be the number of different vectors $\mathbf{c}_{\bar{\beta}}$, $\bar{\beta} = 1, \dots, \bar{\Delta}$, then the external approximation $D^+(t_f, t_0, \mathbf{x}(t_0))$ of the set $D(t_f, t_0, \mathbf{x}(t_0))$ is a polyhedron whose faces lie on the corresponding hyperplanes, the internal approximation $D^-(t_f, t_0, \mathbf{x}(t_0))$ of $D(t_f, t_0, \mathbf{x}(t_0))$ is a polyhedron whose vertices are the points $\mathbf{x}_{\bar{\beta}}^*(t_f)$, i.e., $D^-(t_f, t_0, \mathbf{x}(t_0)) = \text{Co}(\mathbf{x}_1(t_f), \dots, \mathbf{x}_{\bar{\Delta}}(t_f))$. The bigger $\bar{\Delta}$, the better approximation of the attainability set $D(t_f, t_0, \mathbf{x}(t_0))$ can be obtained. It can be proved (Okhtilev et al.,2006, Ivanov and Sokolov 2010, Ivanov et al., 2010) that the value $\bar{\Delta}$ is defined by the total

number of possible interruptions for AMO CS interaction operations over a given time period (t_0, t) .

To obtain D^+ , D^- Krylov and Chernousko's method was used (Chernousko F.L. 1994). Instead of the vector \mathbf{c} the vector $\boldsymbol{\Psi}(t_0)$ of conjugate variables is to be varied.

Besides the general dynamic model of AMO CS functioning (the model M) its aggregated variants can be used for the attainability-set construction. Let us exemplify this approach via the models M_0 , M_k . Interaction operations of the object B_j will be regarded as one aggregated operation, the channels $C_\lambda^{(j)}$ will be replaced by one general channel $C^{(j)}$. Besides, we prescribe $\theta_{i\alpha j\lambda}=1 \forall i, \alpha, j, \lambda$ and allow the interruptions of operations. So the aggregated models of object's IO and channels can be stated as follows:

$$\dot{\tilde{x}}_i^{(o)} = \sum_{j=1}^m \varepsilon_{ij}(t) \tilde{u}_{ij}^{(o)}, \quad (3)$$

$$\dot{\tilde{x}}_{ij}^{(k)} = \sum_{\substack{l=1 \\ l \neq i}}^m \tilde{u}_{lj}^{(k)} \frac{h_{li}^{(j)} - \tilde{x}_{ij}^{(k)}}{\tilde{x}_{ij}^{(k)}} \gamma_- \left(\tilde{x}_{ij}^{(k)} \right), \quad (4)$$

where $\tilde{x}_i^{(o)} = \sum_{\alpha=1}^{s_i} x_{i\alpha}^{(o)}$, $\tilde{u}_{ij}^{(o)} = \sum_{\alpha=1}^{s_i} u_{i\alpha j}^{(o)}$ are the

aggregating functions. The classes $\tilde{K}_\sigma^{(o)}$, $\tilde{K}_\sigma^{(k)}$ of allowable control inputs are defined as follows:

$$\tilde{K}_\sigma^{(o)} = \left\{ \tilde{U}_\sigma^{(o)} = \left\| \tilde{u}_{ij}^{(o)} \right\| \left\| \sum_{i=1}^m \tilde{u}_{ij}^{(o)} \leq 1, \right. \right. \quad (5)$$

$$\left. \sum_{j=1}^m \tilde{u}_{ij}^{(o)} \leq 1, \tilde{u}_{ij}^{(o)} \tilde{x}_{ij}^{(o)} = 0, \tilde{u}_{ij}^{(o)} \in \{0,1\}; \tilde{s}_\sigma^{(o)} \right\},$$

$$\tilde{K}_\sigma^{(k)} = \left\{ \tilde{U}_\sigma^{(k)} = \left\| \tilde{u}_{ij}^{(k)} \right\| \left\| \sum_{i=1}^m \tilde{u}_{ij}^{(k)} \leq 1, \right. \quad (6)$$

$$\left. \sum_{j=1}^m \tilde{u}_{ij}^{(k)} \leq 1, \tilde{u}_{ij}^{(k)} \in \{0,1\}; \tilde{s}_\sigma^{(k)} \right\},$$

where $\tilde{s}_\sigma^{(o)}$, $\tilde{s}_\sigma^{(k)}$ are function-theoretic constraints imposed on the classes of allowable controls.

We assume that the control inputs are piecewise continuous functions. We introduce vector

$$\tilde{\mathbf{x}}^{(o)} = \left\| \tilde{x}_1^{(o)}, \dots, \tilde{x}_m^{(o)} \right\|^T \quad \text{and} \quad \text{vector}$$

$$\tilde{\mathbf{x}}^{(k)} = \left\| \tilde{x}_1^{(k)}, \dots, \tilde{x}_m^{(k)} \right\|^T. \quad \text{Let} \quad \tilde{\mathbf{x}}^{(o)}(t_0) = 0,$$

$\tilde{\mathbf{x}}^{(k)}(t_0) = \tilde{\mathbf{x}}_0^{(k)}$. Then the attainability set in the state space of the dynamic system (3)–(4) can be obtained as follows:

$$\tilde{D}_{(o,k)} = \left\{ \tilde{\mathbf{x}} \left| \tilde{x}_i^{(o)} = \int_{t_0}^{t_f} \sum_{j=1}^m \varepsilon_{ij}(\tau) \tilde{u}_{ij}^{(o)}(\tau) d\tau, \right. \right. \quad (7)$$

$$\left. \tilde{U}_\sigma^{(o)} \in \tilde{K}_\sigma^{(o)}, \right. \\ \left. \tilde{x}_{ij}^{(k)} = \int_{t_0}^{t_f} \sum_{l=1}^m \tilde{q}_{lj}(\tau) \tilde{u}_{lj}^{(k)}(\tau) d\tau, \tilde{U}_\sigma^{(k)} \in \tilde{K}_\sigma^{(k)} \right\},$$

$$\text{where } \mathbf{x} = \left\| (\tilde{x}^{(o)})^T (\tilde{x}^{(k)})^T \right\|^T, \tilde{q}_{lj} = \frac{h_{li}^{(j)} - \tilde{x}_{ij}^{(k)}}{\tilde{x}_{ij}^{(k)}} \gamma_- \left(\tilde{x}_{ij}^{(k)} \right).$$

The following theorem [20] expresses characteristics of the attainability set.

Theorem 1. Let the functions $\varepsilon_{ij}(t)$ be nonnegative bounded functions having at most denumerable points of discontinuity, let the classes of allowable controls be defined by (5), (6), then the attainability set $\tilde{D}_{(o,k)}$

meets the following conditions:

a) It is bounded, closed, and convex. It lies in the nonnegative orthant of the space $\tilde{X} = \mathbf{R}^{(m+mm)}$;

$$\text{b) } D_{(o,k)}^- \subseteq \tilde{D}_{(o,k)} \subseteq D_{(o,k)}^+, \quad (8)$$

Here

$$\tilde{D}_{(o,k)}^- = \left\{ \tilde{\mathbf{x}} \left| 0 \leq \tilde{x}_i^{(o)} \leq \bar{\xi}_i \tilde{x}_i^{(o)}, \right. \right. \quad (9)$$

$$0 \leq \tilde{x}_{ij}^{(k)} \leq \bar{\chi}_i \varphi_{ij}^{(k)}, \bar{\xi}_i \geq 0,$$

$$\left. \sum_{i=1}^m \bar{\xi}_i = 1; 0 \leq \bar{\chi}_i \leq 1 \right\},$$

$$\tilde{D}_{(o,k)}^+ = \left\{ \tilde{\mathbf{x}} \left| 0 \leq \tilde{x}_i^{(o)} \leq \tilde{x}_i^{(o)}, \right. \right. \quad (10)$$

$$0 \leq \tilde{x}_{ij}^{(k)} \leq \bar{\chi}_i \varphi_{ij}^{(k)}, 0 \leq \bar{\chi}_i \leq 1 \Big\},$$

where $\tilde{x}_i^{(o)} = \int_{t_0}^{t_f} \left[\max_{j=1, \dots, m} \varepsilon_{ij}(\tau) d\tau \right]$ under the conditions

$$x_{ij}^{(k)} \equiv 0 \forall t, \forall i, \varphi_{ij}^{(k)} = \max_{j'=1, \dots, m} \{h_{li}^{j'}\} \forall j.$$

The theorem is of high importance for the preliminary analysis of AMO CS control processes, as the calculation of the values $\tilde{x}_i^{(o)}$, $\varphi_{ij}^{(k)}$ is rather simple,

while the sets $D_{(o,k)}^-$, $D_{(o,k)}^+$ let, in many cases, verify the end conditions and calculate the range of variation for the measures of AMO CS ITA.

The sets D , $D^{(+)}$, $D^{(-)}$ and their images in the criteria space can be represented in a graphic form by Cartesian display.

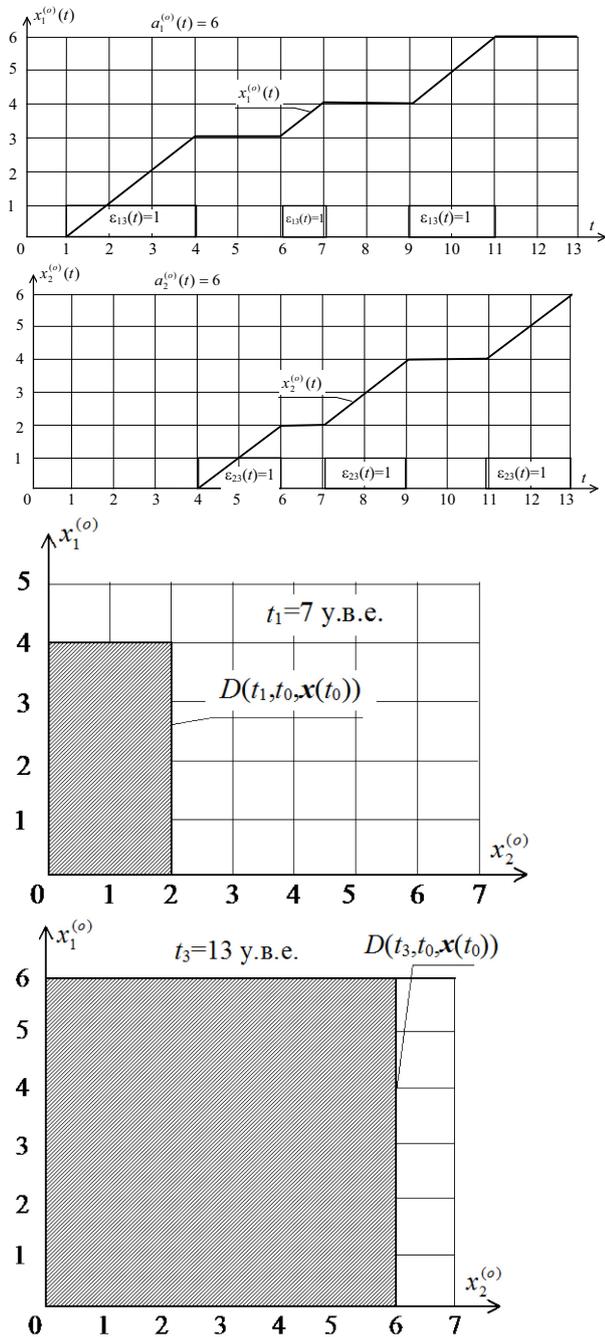
Fig.1-3 illustrate different representation variants for attainability sets. So they show three service situations for the object B_3 interaction with the objects B_1 и B_2 (see the expressions (3), (5)).

The first service situation (see Figure 1) demonstrates the absence of conflicts (interaction zones (IZ) do not intersect).

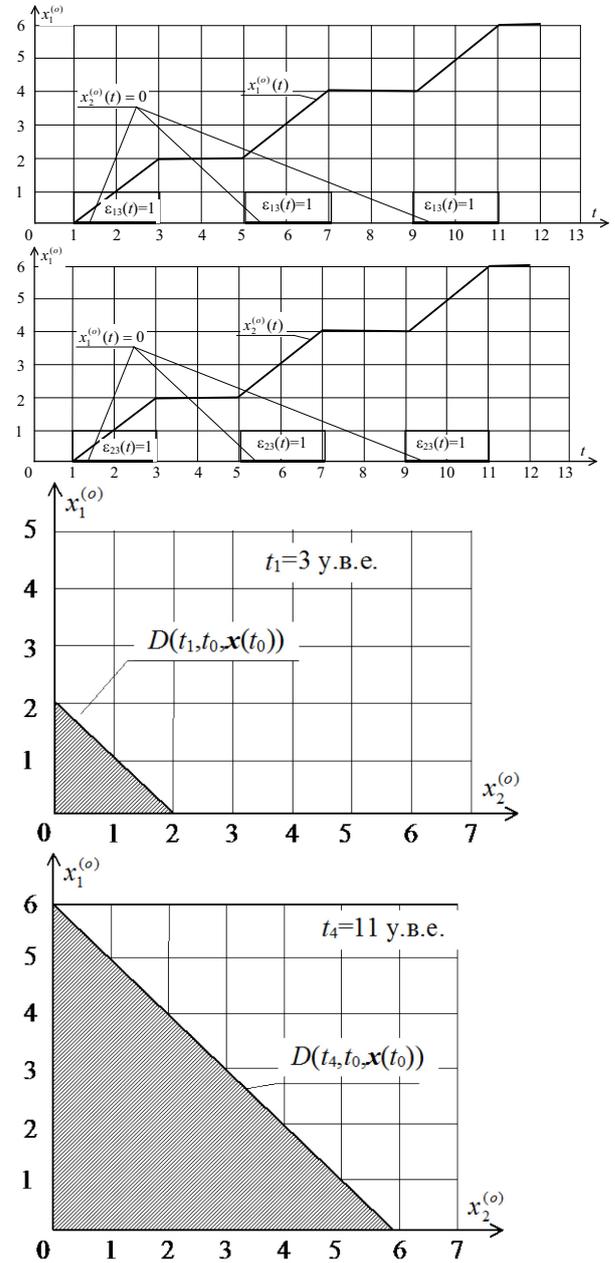
The second service situation (see Figure 2) shows the whole intersection of IZ and maximal conflicts.

The third service situation (see Figure 3) is intermediate and shows the partial intersection of IZ.

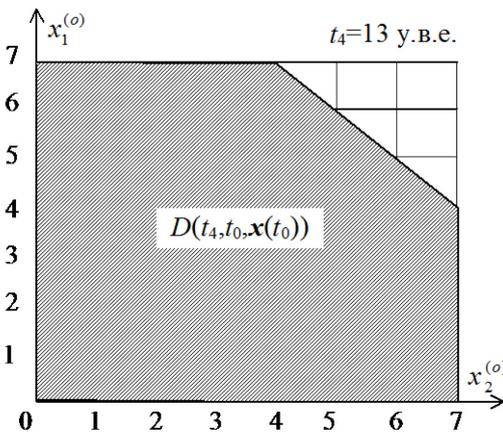
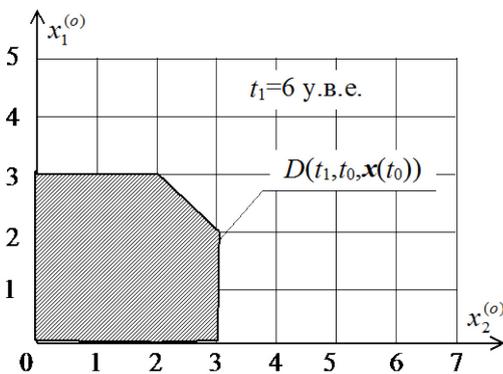
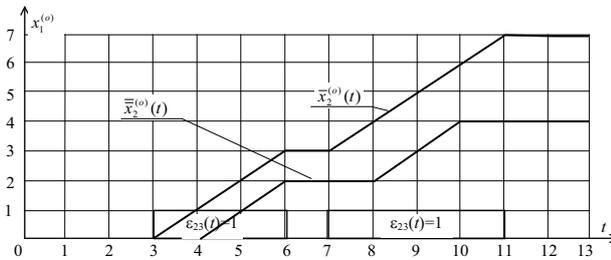
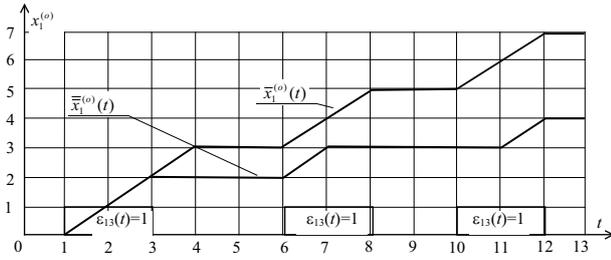
If the number of interacting pairs grows, then several Cartesian systems or polar diagrams may be used.



Figures 1: The first service situation demonstrates the absence of conflicts in AMO CS



Figures 2: The second service situation shows maximal conflicts in AMO CS.



Figures 3: The third service situation is intermediate and shows the partial intersection of interaction zones

CONCLUSION

An attainability set (AS) is a fundamental characteristic of any dynamic system (in our case – AMO CS). The AS approach determines the range of execution policies in the presence of disturbances over which the system can be guaranteed to meet certain goals. An AS in the state space depicts the possible states of AMO CS schedule to variations of the model parameters (e.g.,

different capacities and processing times (Ivanov et al. 2010b). In order to interconnect the schedule execution and the performance analysis to the AS in the state space, an AS in the performance space has to be constructed.

Besides, if an AS we know it becomes possible to analyse the dependence between the AMO CS scheduling results subject to the schedule performance and the start and end states. In other words, it becomes possible to define the area in which permissible solutions (e.g., schedules) are included. On the other hand, an AS analysis may show that, with the given resources and at the given time horizon, it is impossible to achieve the required output performance; hence, additional resources should be introduced or the supply cycle shall be expanded (here, the AS approach is similar to goal programming). Limitations of using AS are their dimensionality. However, in most cases, it is possible to approximate AS, e.g., to a rectangular form while estimating the outcomes at four points of an AS.

ACKNOWLEDGEMENTS

The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277), grant 074-U01 (ITMO University), project 6.1.1 (Peter the Great St.Petersburg Polytechnic University) supported by Government of Russian Federation, Program STC of Union State “Monitoring-SG” (project 1.4.1-1), State research 0073–2014–0009, 0073–2015–0007.

REFERENCES

- Chauhan, S.S, Gordon, V., Proth, J.-M. (2007). Scheduling in supply chain environment. *Europe-an Journal of Operational Research*, 183(3), 961-970.
- Chernousko F.L (1994) *State Estimation of Dynamic Systems*. SRC Press, Boca Raton, Florida
- Clarke FH, Ledyaev Yu S, Stern RJ, Wolenskii PR (1995) Qualitative properties of trajectories of control systems: a survey. *J Dyn Control Syst* 1:1–48.
- Gubarev, V.A., Zakharov, V.V., Kovalenko, A.N. (1988). *Introduction to systems analysis*. LGU, Leningrad.
- Ivanov, D., Sokolov, B. (2010), *Adaptive Supply Chain Management*, Springer, London et al.
- Ivanov, D., Sokolov, B., Kaeschel, J. (2010). A multi-structural framework for adaptive supply chain planning and operations with structure dynamics considerations. *European Journal of Operational Research*, 200(2), 409-420.
- Kalinin, V.N., Sokolov, B.V. (1985). Optimal planning of the process of interaction of moving operating objects. *International Journal of Difference Equations*, 21(5), 502-506.
- Kalinin, V.N., Sokolov, B.V. (1987). A dynamic model and an optimal scheduling algorithm for activities with bans of interrupts. *Automation and Remote Control*, 48(1-2), 88-94.
- Khmel'nitsky, E., Kogan K., & Maimom, O. (1997). Maximum principle-based methods for production scheduling with partially sequence-dependent setups.

International Journal of Production Research, 35(10), 2701–2712.

Klir G. (2005). *Uncertainty and Information: Foundations of Generalized Information Theory*, John Wiley, Hoboken, NJ.

Okhtilev, M.Y., Sokolov, B.V., Yusupov, R.M. (2006). *Intellectual Technologies of Monitoring and Controlling the Dynamics of Complex Technical Objects*. Moskva: Nauka, p. 409.

Vieira, G.E., Herrmann, J.W., and Lin, E. (2000). Predicting the performance of rescheduling strategies for parallel machine systems. *Journal of manufacturing Systems*, 19(4), 256-266.

AUTHOR BIOGRAPHIES

BORIS SOKOLOV is a deputy director at the Russian Academy of Science, Saint Petersburg Institute of Informatics and Automation. Professor Sokolov is the author of a new scientific lead: optimal control theory for structure dynamics of complex systems. Research interests: basic and applied research in mathematical modelling and mathematical methods in scientific research, optimal control theory, mathematical models and methods of support and decision making in complex organization-technical systems under uncertainties and multicriteria. He is the author and co-author of 9 books on systems and control theory and of more than 450 scientific papers. Professor B. Sokolov supervised more over 90 research and engineering projects. His e-mail address is sokolov_boris@inbox.ru and his Web-page can be found at <http://litsam.ru>.

DMITRY IVANOV is a professor at Department of Business and Economics at the Berlin School of Economics and Law and Chair of the German-Russian Coordination Office for Logistic. He studied production management and engineering (2000). In 2002, he graduated in Saint Petersburg as a Ph.D. in Economics on the topic of operative supply chain planning and control in virtual enterprises. In 2006, he received the

Dr.rer.pol. degree at the Chemnitz University of Technology. He is an author of 6 scientific books and more than 70 papers published in international and national journals, books and conference proceedings. Since 2001, he has been involved in research and industry projects on supply chain management and virtual enterprises. Dr. Ivanov received a German Chancellor Scholarship Award in 2005. His e-mail address is: idm@hrz.tu-chemnitz.de.

VLADIMIR KALININ Dr. Sc.in Technical Science, Prof., Honored Scientists of the Russian Federation; The full member of the Russian academy of astronautics of a name K.E. Tsiolkovsky, Professor of Military-space academy. Research interests – the theory of system researches, space cybernetics and computer science, the theory of optimum control of the dynamic systems, the automated control systems, preparation of the engineering staff and new information-didactic technologies in higher education. Number of scientific publications – 170. His e-mail address is kvn.112@mail.ru.

SERGEY NEMYKIN. In 1988 graduated from the Moscow aviation institute named after S.Ordzhonikidze as "electrical and mechanical engineer". From 1988 to 2013 – the engineer, the head of group, the chief of sector, the chief of office, the first deputy head of skilled design office Federal State Unitary Enterprise Scientific and Production Association named after S.A. Lavochkina. From 2013 to 2015 –Federal State Unitary Enterprise Central Research Institute of Chemistry and Mechanics. Since April 6, 2015 – the general designer of The Design Bureau “Arsenal” named after M.V.Frunze. Author of 15 scientific works. His e-mail address is kbarsenal@kbarsenail.ru.

EMERGENCY PREDICTION IN ELECTRIC UTILITIES: A CASE STUDY FROM SOUTH BRAZIL

Iochane Guimarães, Vinicius Jacques Garcia and
Daniel Pinheiro Bernardon
Federal University of Santa Maria UFSM, Brazil.
Email: viniciusjg@ufsm.br

Julio Fonini
AES Sul - Power Utility, Brazil.

KEYWORDS

Emergency orders, forecasting, dynamic vehicle routing.

ABSTRACT

This work proposes a methodology to predict emergency requests on electric distribution utilities, based on the historical data in order to promote a further pro-active routing by reducing a dynamic vehicle routing in its static form. The proposed methodology aims to minimize the average service time for emergency services, from the consideration of typical aspects that exert influence on the service time required: the day of the week, time of the day and the geographical location where services are requested. A case study from actual data is presented to show how the methodology can be applied.

INTRODUCTION

In the electric power utilities sector, there is a set of orders to be attended by a set of teams. The set of orders includes two subsets: the subset of commercial orders and the subset of emergency services. The former is known in advance and the latter is only known when vehicles are proceeding with their routes (Garcia et al. 2014).

The attendance of emergency services in electric utilities corresponds to a major and a high impact task related to the Network Operation Center. Not only by assuring security procedures but also by promoting efficiency and effectiveness on the services provided, accomplishing these random requests is a challenge especially with the occurrence of extreme climate changings, as in the South Brazil.

Since the Network Operation Center also manages service requests of non-emergency character, a general concern involves how to plan maintenance crew routes from the assumption that there will be unknown requests that will have to be attended immediately. The optimization problem involved in the routing planning corresponds to the vehicle routing problem (Toth and Vigo 2001), a well-studied problem with a large range of contributions (Eksioglu et al. 2009).

When assuming a scenario to handle emergency requests, the vehicle routing problem takes its dynamic form (Psaraftis 1995; Pillac et al. 2013). By knowing

gradually these emergency requests in such a way that vehicles are following their pre-established routes, the question that arises from this context is how much time to wait in order to reprogramming their paths: (Larsen et al. 2002) suggest that the “degree of dynamism” may be considered. Another approach may be to predict these stochastic requests in order to obtain a static vehicle routing problem, which is the proposal of this paper.

This work presents a methodology to predict emergency services in an attempt to answer the following questions when a pro-active routing is assumed:

1. Which is the period of time over the time horizon considered that will be emergency requests?
2. Which are the locations where the random requests will occur?
3. How much time each emergency request predicted will involve as service time?

By answering the first question, the proposed methodology furnishes a time window to be considered in the static vehicle routing. The answer to the second question corresponds to the geographic consideration of dummy nodes in the vehicle routing problem. And finally, answering the last question one defines the service time to the dummy nodes previously created.

It is worth noting that the contribution of this work is only part of a more sophisticated dispatch system, as pointed out by (Weintraub et al. 1999). The following sections detail the proposed methodology, the case study developed from actual data and finally the final remarks.

METHODOLOGY

Planning routes to maintenance crews is part of a dispatching process in electric distribution utilities, from the consideration of two main kinds of orders, which can be described as (Garcia et al. 2014):

- a) The commercial orders: known in advance and are typically created from customer requests;
- b) The emergency orders: these orders are not known a priori and can occur at any moment.

The dispatch of commercial orders is performed by the method proposed by (Garcia et al. 2014), comprising the well-know vehicle routing problem in its static form. From the business processes usually employed by the

electric distribution utilities, the same maintenance crew that serves commercial orders should also attend emergency requests, thus qualifying a multitasking character and involving a partially dynamic vehicle routing problem as described by (Larsen et al. 2002).

Figure 1 illustrates a hypothetical routing solution to commercial orders: route 1, starting at node 1 and visiting nodes 3, 5 and 7; route 2, starting at node 2 and visiting nodes 8, 6 and 4. Since this routing problem involves on site service time due to the nature of the problem of attending service orders (Garcia et al. 2014), the arrival time at each node should be calculated. For this example, they are presented on Table 1 in minutes in the last column, as well as the service time also presented in minutes.

When considering partially dynamic scenarios, a certain number of emergency orders come up, two as the example of Figure 2 (emergency orders 9 and 10). The question that may arise is the moment when this perturbation occurs: if we consider the occurrence of both orders at instant $t=0$, a possible solution may be that one presented in Figure 3. However, if this occurrence is on instant of time $t=27$, only after completing services 3 and 8 that both 9 and 10 come up, making the most preemptive behavior as that presented in the solution of Figure 4.

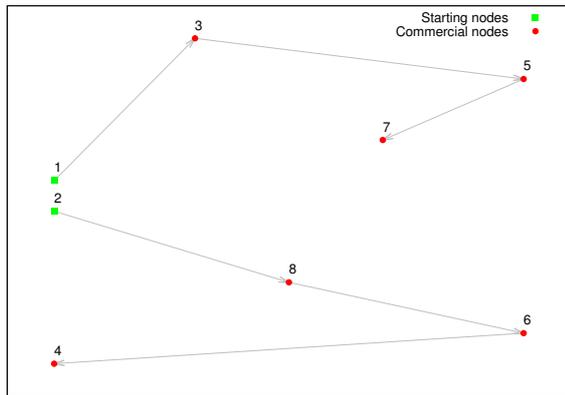


Figure 1: A hypothetical routing solution to commercial orders.

Table 1: Arrival times for the example of Figure 1.

Route	Node	Service time	Arrival time
1	1	0	0
1	3	10	15.23
1	5	3	39.79
1	7	55	60.27
2	2	0	0
2	8	6	12.20
2	6	3	29.38
2	4	30	55.61

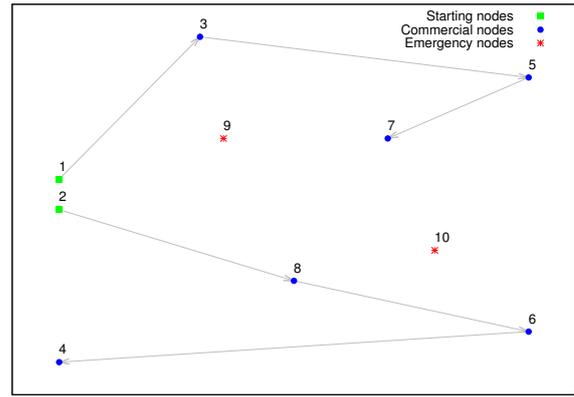


Figure 2: Hypothetical instance of emergency dispatch problem.

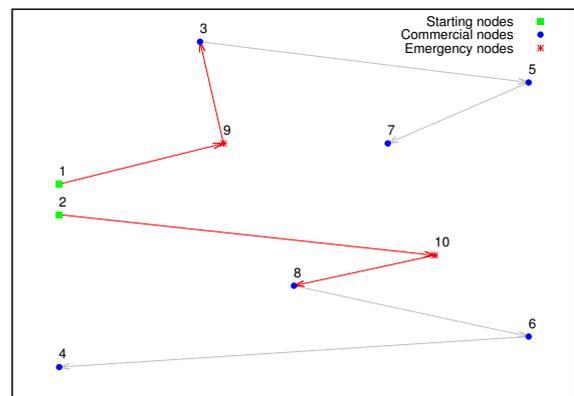


Figure 3: Solution for instance of Figure 2 ($t=0$).

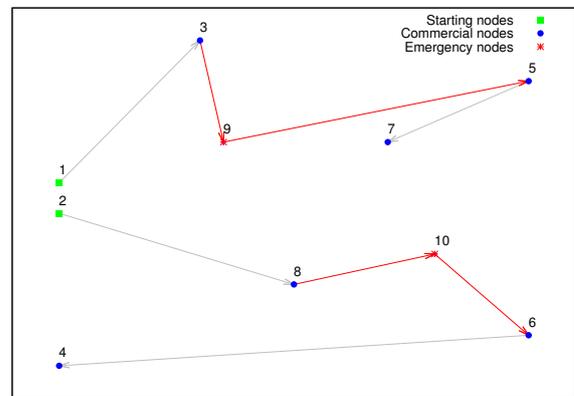


Figure 4: Solution for instance of Figure 2 ($t=27$).

The main inspiration for the proposed approach to handle this partially dynamic vehicle routing problem comes from the recent work of (Ferrucci et al. 2013), which describes a pro-active real-time control approach for dynamic vehicle routing problems. Following the same approach of Ferrucci et al. that considers historical request information, future emergency requests are predicted without assuming any prescribed probability distributions.

This work presents the following attributes as those fundamental to be considered in carrying out the emergency requests: (i) the location of the service, i.e. where is the customer order demand; (ii) time of occurrence; and (iii) service time. At the same time, it is assumed that the management of services to handle the assignment and routing decisions related to the maintenance crews involves observance of route time, since these crews have strict workload which may not be violated.

The definition of which variables significantly influence the occurrence of an emergency service order has been performed from the correlation analysis between the following variables: (1) Cause of the emergency event; (2) Latitude; (3) Longitude; (4) Year; (5) Month; (6) Day of the month; (7) Weekday; (8) Time of day; and (9) Service time. Afterwards, random variables and their domains (continuous or discrete) are defined. It is known that each random variable is quantified by a probability density function, so to identify the distribution of each variable the following steps were followed as (Taha 2007):

1. Summary of the raw data in the form of a suitable frequency histogram function to determine the empirical probability density associated with the random variable (day, time and geographical location);
2. Analysis of the dispersion of the number of service time hours by each geographical area considered;
3. From the definition of empirical Probability Density Function that every random variable (X) belongs, are calculated the expected value $E(x)$ and standard deviation (σ);
4. Calculation of probabilities from frequency distribution ranges for service time considering day of week, time of day and the discretization of the geographical area assumed.

This method of prediction added the service time for every day of the week and every hour of day, with an estimated probability of occurrence in a particular geographic area.

Following these steps, one can obtain the dummy nodes, with their corresponding geographic locations and service time.

From a broader perspective, one can be easily situated on how to use this procedure in a more sophisticated dispatch system following the flowchart of Figure 5. At the beginning, all the information about teams is available, called "Team data", like workday hours, average speed and location. At the same time, all the information about the orders is also available from the database system, depicted as "node data", which means to capture location, service time and priority of those orders assumed as commercial ones. From the historical data, the procedure namely "dummy node prediction"

corresponds to the methodology described by steps 1-4, in such a way that these nodes can be assumed a-priori by the dispatch system, thus performing a static vehicle routing. By executing "vehicle routing algorithm", one obtains the planned routes to all teams considered and including not only the commercial orders but also an "estimation" of possible occurrence of "emergency nodes", namely "dummy nodes", as previously depicted in Figure 1-Figure 4.

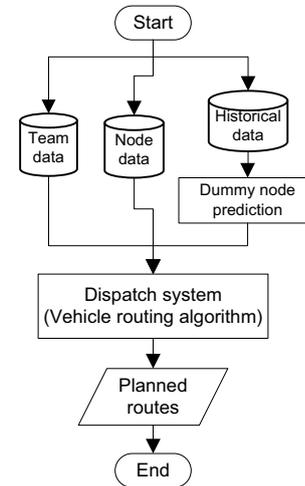


Figure 5: The proposed methodology in the whole dispatch system.

The next section presents the case study developed to show how to apply these four steps.

CASE STUDY

The case study developed for testing and validating the proposed methodology is based on the observation 27989 occurrences of emergency services in an electric distribution utility, comprising a horizon of 392 days.

After accumulating these 27989 occurrences in each of 24 hours of day, the sample size is equal to 9408 units. Figure 6 depicted the service time in hours for each one of these 9408 units. After reducing the relative standard deviation from 102% to 79%, by removing approximately 3.6% of the sample, one obtains a reduced sample as shown in Figure 7.

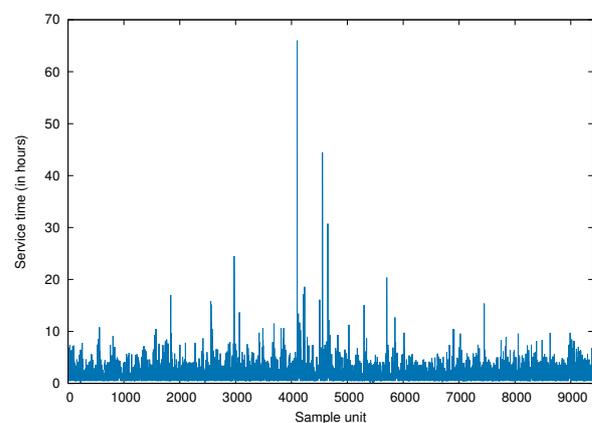


Figure 6: Service time for raw data.

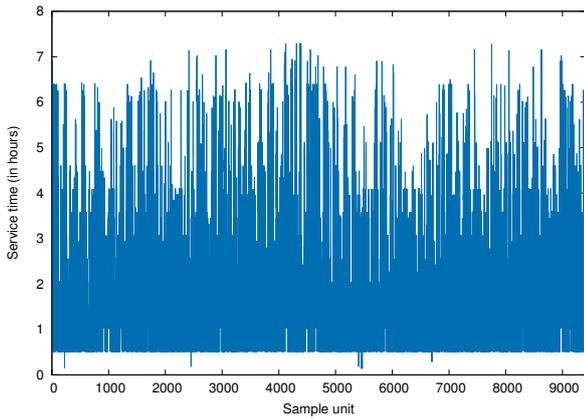


Figure 7: Service time for the reduced sample.

Next, it is necessary to carefully observe how data behave as the random variables assumed on the day of the week, the time of day and the geographical distribution. Figure 8, Figure 9 and Figure 10 analyze the reduced sample for the day of the week, the time of day and the geographical distribution, respectively. The latitude and longitude axis of Figure 10 are discretized according to the number of each box defined by following a pre-established size 16 km^2 . This size is justified from the moment that routing for planning purposes, displacements within each box are disregarded.

From the analysis of Figure 8-Figure 10, one can conclude the following:

- Business days have similar behavior; Sundays and Saturdays require an individual analysis;
- In fact the time of day must be assumed in particular for every day of the week considered;
- Discretization of the geographical area in boxes of 16 km^2 ($4 \text{ km} \times 4 \text{ km}$) allows the consideration about high demand regions.

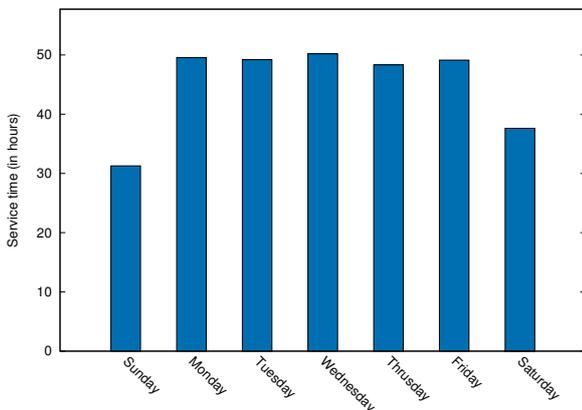


Figure 8: Analysis for the variation on “day of the week” variable.

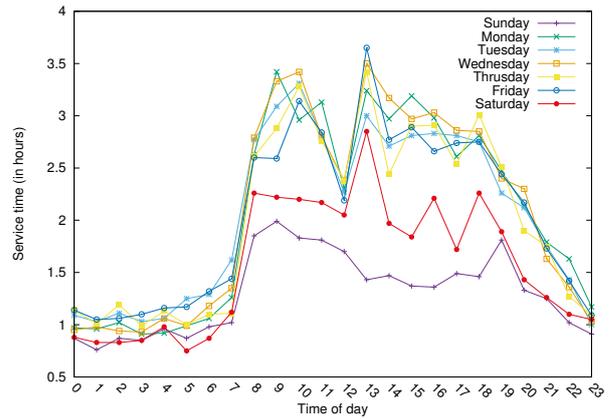


Figure 9: Analysis for the variation on “time of day” variable.

The next step corresponds to an attempt to answer the questions pointed out in the first section of this work: (1) Which is the period of time over the time horizon considered that will be emergency requests? According to Figure 9, the period between 7 and 20 hours of any business day is the most relevant to be considered; (2) Which are the locations where the random requests will occur? According to Figure 10, only a small subset of the potential locations are the most relevant with regard to the service time required; (3) How much time each emergency request predicted will involve as service time? The service time of each box of Figure 10 may help to evaluate how many working hours will be needed.

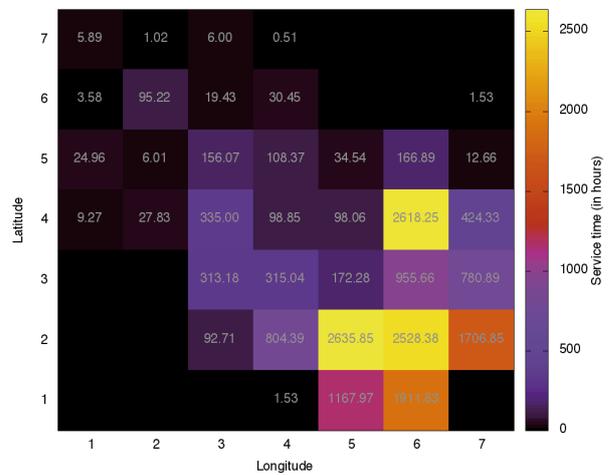


Figure 10: Analysis for the variation on “geographical distribution” variable.

The next step refers to the measurement of these quantities and the sample separation strategies in defined quantities. Considering that we are trying to predict the level of service time required to a period of a certain Tuesday, between 8 am and 12 pm, one must find the historical data for all occurrences of a Tuesday, between 8:12 in the morning, as Figure 11 summarizes. After that, occurrences need to be stratified in the geographical area according to the set division: since time of day is very important, Figure 12-Figure 16

includes the service time of each hour between 8 and 12 in the morning over the discretization of Figure 10.

With this result, each hour (8:12) has the its service time demand stratified within 49 boxes, and one may conclude that the area with the highest demand is not the same when comparing the hours.

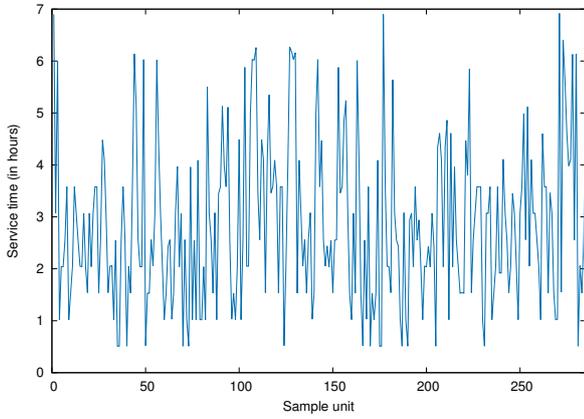


Figure 11: Service time for the Tuesday, between 8 am and 12 pm.

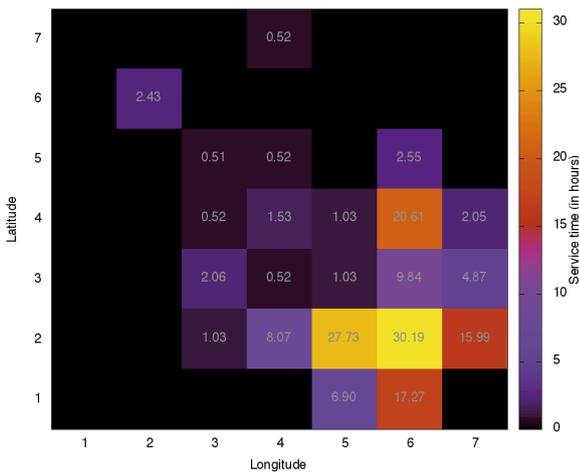


Figure 12: Historical service time on Tuesday, 8 am.

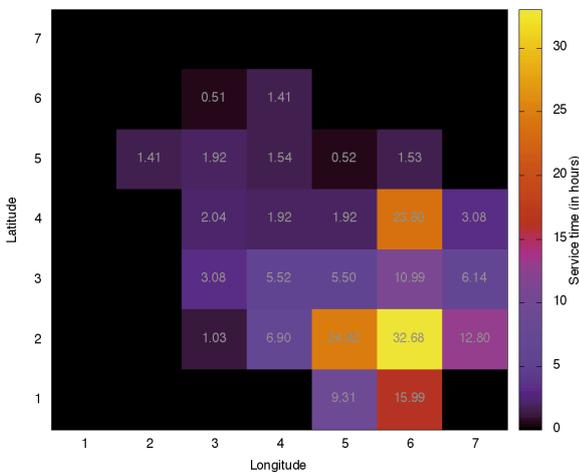


Figure 13: Historical service time on Tuesday, 9 am.

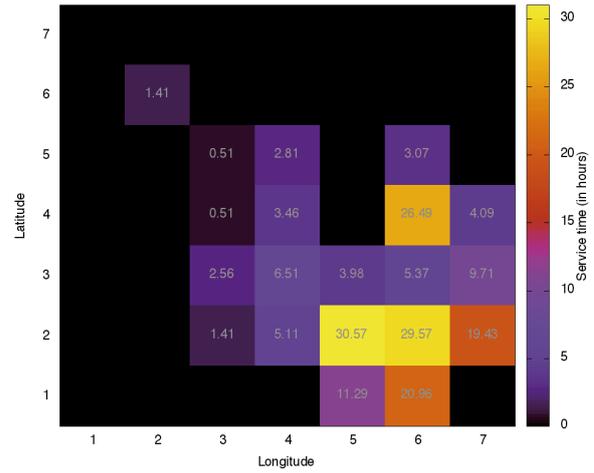


Figure 14: Historical service time on Tuesday, 10 am.

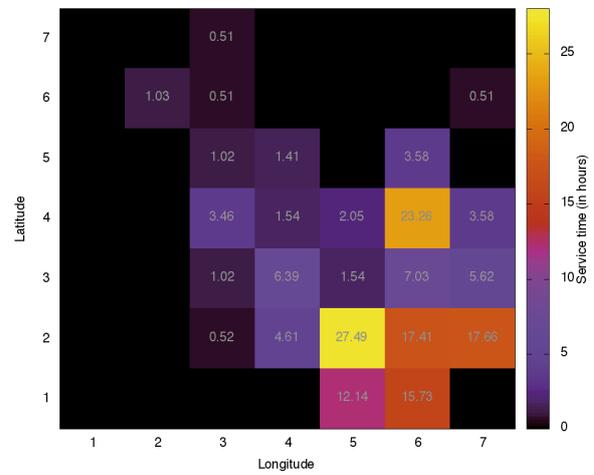


Figure 15: Historical service time on Tuesday, 11 am.

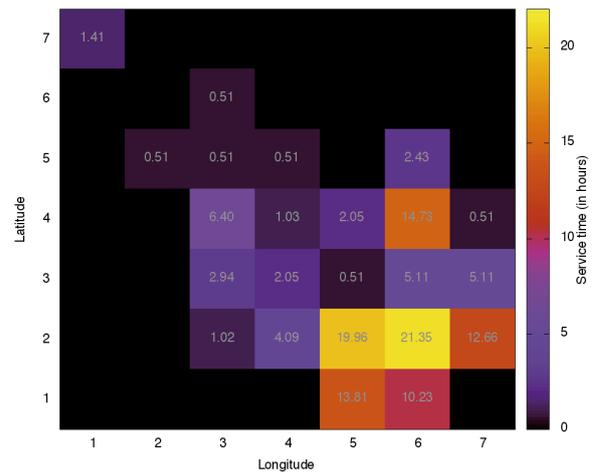


Figure 16: Historical service time on Tuesday, 12 pm.

Now it is time to define the level of service time in each area selected, for every hour considered: 8 am – 12 pm. According to (Taha 2007), this calculation may be done by assuming a random variable with empirical distribution and obtaining the expected value of each one of these variables: each variable corresponds to a box previously selected. First it is defined the histogram

information of each box for each hour, for instance, box (2;6) of Tuesday, 8 am, Figure 12. Table 2 summarizes these results, resulting in a expected value of 0.56253 hours of service time.

Table 2: Histogram information about the service time for box (2;6) of Tuesday, 8 am.

# Interval	Range	Observed frequency	Relative frequency	Cumulative frequency
1	[0-0.51090]	48	0.92308	0.92308
2	(0.51090-0.73465]	0	0	0.92308
3	(0.73465-0.95840]	0	0	0.92308
4	(0.95840-1.18215]	4	0.07692	1.00000

These calculations are proceed for all boxes in each hour, thus resulting a new map, this turn showing the expected value of service time for each box, as pointed out in Figure 17 for Tuesday, 8 am.

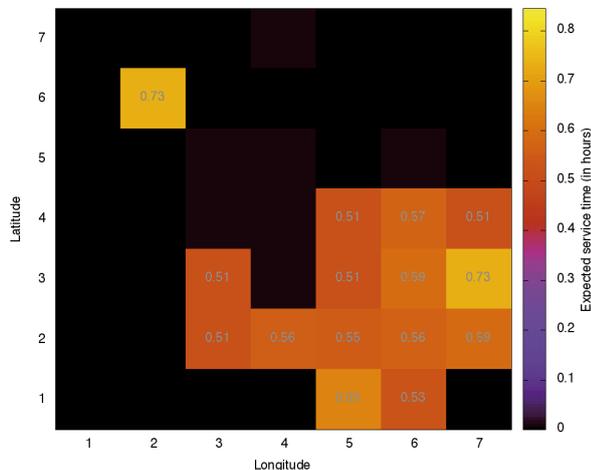


Figure 17: Expected service time for Tuesday, 8 am.

FINAL REMARKS

This paper has presented a methodology to predict emergency requests in electric distribution utilities, based on historical data in order to allow a further proactive routing, which is understood as a way of reducing a dynamic vehicle routing in its static form.

Even considering that this approach has a large number of open questions about the effectiveness, the simplicity involved turn easy its consideration on actual scenarios at least to obtain a estimation about lower bounds on route efficiency over a certain planning horizon.

Another favorable point refers to the low response time and the effort required to be self-adaptive over the time.

ACKNOWLEDGEMENTS

The authors would like to thank the AES SUL Distribuidora Gaúcha de Energia SA for financial support provided to the project “Planejamento dinâmico de Operações”.

REFERENCES

- Eksioglu, B; A. V. Vural and A. Reisman. 2009. “The vehicle routing problem: A taxonomic review”. *Computers & Industrial Engineering*, vol. 57, no. 4, pp. 1472–1483.
- Elsayed, E. A.; T.O. Boucher. 1994. “Analysis and Control of Production Systems”. New Jersey: Prentice Hall.
- Ferrucci, F.; S. Bock, and M. Gendreau. 2013. “A pro-active real-time control approach for dynamic vehicle routing problems dealing with the delivery of urgent goods”. *European Journal of Operational Research*, vol. 225, no. 1, pp. 130–141.
- Garcia, V.J.; D.P. Bernardon; A. Abaide and J. Fonini, J. 2014. “Multi-criteria approach for emergency service orders in electric utilities”. In *Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014*. pp. 676.
- Larsen, A.; O. Madsen and M. Solomon. 2002. “Partially dynamic vehicle routing: models and algorithms”. *Journal of the Operational Research Society*, 53:637–646.
- Pillac, V.; M. Gendreau; C. Guéret and A.L. Medaglia. 2013. “A review of dynamic vehicle routing problems”. *European Journal of Operational Research*, vol. 225, no. 1, pp. 1–11.
- Psaraftis, H. N. 1995. “Dynamic vehicle routing : Status and prospects”. *Annals of Operations Research*, vol. 61, pp. 143–164.
- Taha, H. A. 2007. “Operations research: an introduction”. Pearson Prentice Hall, 8th ed.
- Toth, P. and D. Vigo. 2001. “The Vehicle Routing Problem”. *Monographs on Discrete Mathematics and Applications*, SIAM.
- Weintraub, A.; J. Aboud; C. Fernandez; G. Laporte and E. Ramirez. 1999. “An emergency vehicle dispatching system for an electric utility in Chile”. *Journal of the Operational Research Society*, 50, 690–696.

AUTHOR BIOGRAPHIES

IOCHANE GUIMARÃES received his Bachelor's degree on Production Engineering from the Federal University of Pampa, in 2012. Currently, she is enrolled in her master course on Production Engineering at Federal University of Santa Maria, in Brazil.

VINÍCIUS JACQUES GARCIA 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.

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.

Investigation of genetic operators and priority heuristics for simulation based optimization of Multi-Mode Resource Constrained Multi-Project Scheduling Problems (MMRCMPSP)

Mathias Kühn
Taiba Zahid
Michael Völker
Professorship of Logistics
Engineering
TU Dresden, 01062 Dresden,
Germany

Zhugen Zhou
Oliver Rose
Department of Computer Science
University of the Federal Armed
Forces Munich
85577 Neubiberg, Germany

KEYWORDS

Project Scheduling, Meta-Heuristics, Priority Rules, Genetic Algorithm, Assembly Line

ABSTRACT

Solving NP-hard Problems like Multi-Mode Resource Constrained Multi-Project Scheduling Problems (MMRCMPSP) needs efficient search and optimization strategies. The combination of different approaches such as a meta-heuristic (Genetic Algorithm) for the mode assignment and a Heuristic (Priority Rules) for the job selection allows a 2-step solving-process. In this paper, we present such an approach for solving MMRCMPSP implemented with a simulation-based optimization tool. We investigate the influence of specific parameters of the algorithm to figure out which parameters mostly affect the result of MMRCMPSP.

INTRODUCTION

Organizations deal with different sorts of projects subjected to various variables and constraints. The basic principles in project management are the same but they cannot be applied on every organization due to their divergent production layouts and needs. Richard P. Olsen (1971) defined in his article "Can Project Management Be Defined?" project management as "...the application of a collection of tools and techniques to direct the use of diverse resources towards the accomplishment of a unique, complex, one-time task within time, cost, and quality constraints. Each task requires a particular mix of these tools and techniques structured to fit the task environment and life cycle (from conception to completion) of the task."

Resource constrained scheduling is a class of project management problem concerning limited availability of resources which was initially defined by Conway et al. (1967) and later on was further extended by Brucker et al. (2004). The state-of-the-art extension of taxonomy and division can be found in Zahid et al. (2015).

The current paper deals with the extended version of such constrained scheduling problems by means of creating models which are able to reproduce characteristics of real time production floors. It is

known as Multi-Mode Resource Constrained Multi-Project Scheduling Problem (MMRCMPSP).

The two widely known generalizations of modelling multiple modes are (1) an activity with requirement of multiple skills available with various resources (2) change in activity duration time with different quantity of resource. For detailed definitions and variations for MMRCMPSP, we refer to the research provided by Kuster et al. (2007).

In complex industrial environments with a variety of orders and resource limitations, scheduling is an NP-hard problem (Bowman, 1959), even if it is to find a feasible one. This class of problem is characterized as NP hard because of various reasons:

- Integer variables
- Non-convex problems
- Lack of exact solution methodologies

Many scheduling algorithms have been presented in past literature and a few basic techniques can be reviewed in a journal by Kolisch et al. (1995) in which they emphasized the need of computational efficiency for solving large scale problems. With changing customers' requirements and technology, objectives of the companies have been evolved and have multiple dimensions which impose limitations on using traditional search techniques for near optimal results. Solution methodologies for these can be divided into four main types (Brucker, 2004):

Relaxation: The problems are solved by relaxing some parameters.

Approximation: These methods solve the problem close to original one but do not guarantee global optima. These include commonly heuristics methods usually applied on large scale problems.

Expert Systems: These are increasingly gaining popularity for solving NP-hard problems. They usually combine the advantages of enumeration and heuristic methods.

Enumeration: These methods guarantee global optima but are found difficult to apply in case of large variables. Examples include B & B and Dynamic programming.

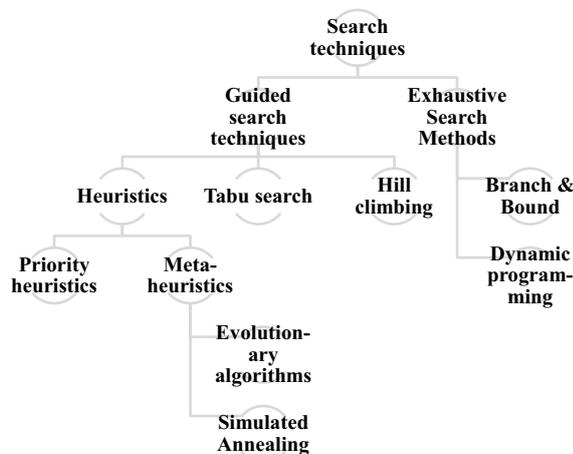
Apart from these solution methodologies, simulation methods and graphical techniques are also being applied

in this area with the main focus of attaining an acceptable optimal schedule in a feasible time for complex production systems. However, studies mostly seem to focus on one aspect of the problem, which is either to find better search algorithms for optimized results or on the development of models.

This paper uses MMRCMPSP model where each resource has various skills required to perform tasks. This way of modelling not only enables decision managers to use various options of shifting activities to a different renewable resource available at the time but also narrates the real time production floor in a better way. On the contrary, it also complicates the problem by increasing the decision variables since apart from the decision of allocating start time for the tasks, assignment of particular resource needs to be done as well. In the next section, review of search techniques in this area has been provided before describing the proposed simulation based optimization strategy.

LITERATURE REVIEW

Due to the number of limitations which came across for using exhaustive search (based on searching the complete search space), scientists were prone towards guided search techniques. The figure below (Zahid, T., 2013) depicts the overview of search techniques for optimization which have been explored so far in the manufacturing industry.



Figures 1: Classification of Search techniques

The research area regarding heuristic solution techniques dominates exact approaches in this area. The basic reason is the problem of complexity faced in RCPSP and its variants while solving large scale problems where exact approaches are unable to find solution with accepted computational time.

The most famous commonly used approach in this area is constructive heuristics which uses priority rules with parallel or series schedule generation schemes. Various types of priority rules have been applied and discussed in literature. These most commonly used greedy priority heuristics are used as a decisive tool to allocate limited resources to tasks.

In a similar study on heuristics (Myszkowski et al. 2013) conducted on new data set developed with the help of Volvo-IT department in Wroclaw. They concluded that resource based priority rule perform the worst and slack based measures were simple in calculations and provide better performance results. In another research (Buddhakulsomsiri and Kim 2007) FIFO was concluded as the best in case of maximizing number of finished products while shortest processing times rule (SPT) performed better with the objective of maximum machine utilization. The comparison of famous priority rules can be seen in an article by Iringova et al. (2012). Another comprehensive study in this regard can be seen in the article by Boctor (1993) who studied state of the art priority rules with respect to different manufacturing layouts. In a research conducted by Kühn et al. (2015), sensitivity analysis was performed to measure the impact of priority rules on various factors of a single performance measure.

With the increase of problem scale and need to extend these optimization techniques to real scale problems, numerous meta-heuristics like simulated annealing (SA), ant colony (ACO), genetic algorithms (GA) and particle swarm optimization (PSO) have been explored to speed up the optimization process. GA is most popular in this regard. GA based decomposition was proposed by Vanhoucke et al. 2013. However, this decomposition was used for optimization of schedule. First the whole schedule was created and then a random part of the schedule was selected for optimization and after achieving the target value, was remerged into the complete schedule. Cheng et al. (2012) used a simulation based GA approach for similar problem with the addition of time based resources.

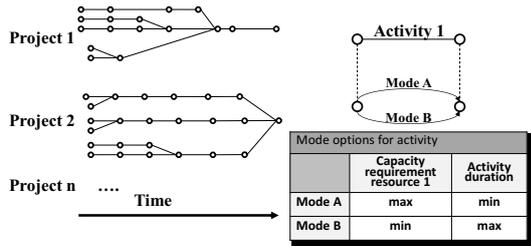
However, GA involves the use of several parameters to search in the solution space with a random distribution such as selection pressure, crossover and mutation rate. In the light of state-of-the-art research, this investigation focuses on the question “How do the GA parameters effect the search direction in correspondence with various priority heuristics?” Among several heuristics, best considered in past researches have been chosen and investigated in combination with GA. As in a conducted study on MMRCMPSP (Kuster, 2007), it was concluded that GA techniques mostly seem to show higher rate of improvements in the first simulation runs which is why good initial solutions are of significant improvement for solution quality. Thus, the present research proposes to investigate priority heuristics and GA in concurrent. In the next section, model and design of experiments would be described in detail for the readers.

MODEL AND PLATFORM

Model

The case study is based on a complex assembly line for the production of printing machines and therefore, attributable to the mechanical and plant engineering and is implemented through a case study. The production system involves different orders considered as separate

projects with individual due dates (Multi-Project-Problem). Every product is an independent project which includes specific product, resource and process information. The individual network of each project is presented as an activity-on-arc network. An activity requires specific qualification/skills for execution, fulfilled by various resources (Modi-Problem). This information is summarized as options in a qualification table (Fig. 2) based on the work of Carl and Angelidis (2015). Following paragraph describes the constraints and variables which were modelled according to the above described production plant requirements.

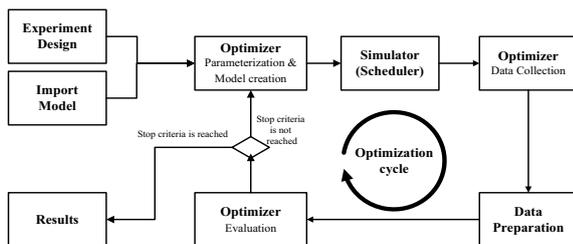


Figures 2: Multi-Mode- and Multi-Project-Problem

Three types of staff resources are considered in this model based on a case study (e.g. electrician, logistic) and each renewable resource is defined via a set of skills/qualifications. In total there are 9 internal workers (assumption: available 24 hours) in addition to the possibility of one subcontracted worker for each resource with increased cost. Equipment (Crane in this case) is also added with possibility of renting another one at an additional cost. The considered scenario includes 10 sales orders with a total of 13 products. The time horizon for production plan is 20 days. Each product has an own due date and cost rate for delay and earliness. The total activities are 2700 which can be performed with a maximum of two modes. There is a range of 126-276 activities to produce for a single product where the activity duration varies between 11 and 400 minutes.

Platform

Simulation based optimization is widely used to get fast optimized solutions in large search spaces such as for the above described problem class. A platform which implements this theory was developed by Angelidis et al. (2013) and is called SBOP (simulation based optimization platform). The optimization cycle is shown in figure 3.



Figures 3: Scheduling with simulation based optimization (Based on Angelidis et al. 2013)

The first step is the transformation of the real system to the manufacturing model. The models describe all necessary production information as previously described in the model part. The optimizer has 3 functions: (1) parameter setting for creation of various simulation models (2) data collection (3) evaluation by calculating the interested key performance indicators. If the stopping criteria are not met, the optimization process continues.

In the later section, optimization objectives and optimizer used for the current simulation would be explained.

Optimizer

In a manufacturing environment, various conflicting objectives are desired for example production cost and production time. For this, production managers need to find the compromise between these conflicting objectives for optimal results. To simultaneously optimize the multi-objective problem, we developed a multi-objective genetic algorithm which (1) utilizes classical Pareto ranking approach to determine the non-dominated solutions (2) introduces methods for diverse population. We will briefly describe the features of this algorithm as follows.

(1) Pareto ranking approach:

It utilizes the Pareto dominance concept (Alvarez-Benitez et al. 2015) to compare the solutions in order to determine the non-dominated solutions in the solution space. In this algorithm we use the Fast Non-dominated Sorting Genetic Algorithm approach referred to Deb et al. 2002.

(2) Customized genetic algorithm to produce diverse population:

(2.1) Chromosome representation (figures 4)

In this study, to represent a simulation model MMRCMPSP, we have used the encoding where each chromosome is made of n genes, where n is the number of the activities.

$$\text{Chromosome} = \underbrace{(\text{gene}_1, \dots, \text{gene}_n)}_{\text{Mode of activity}}$$

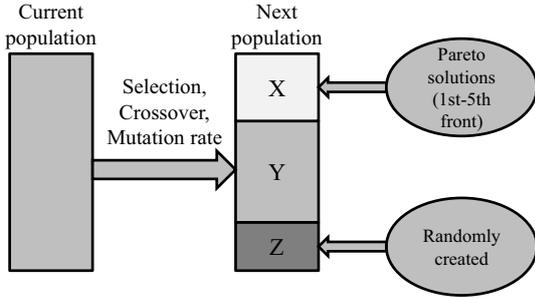
Figures 4: Chromosome representation

(2.2) Population formation (figures 5)

To make a trade-off between convergence speed and diverse population, we have applied a special strategy.

A population consists of three parts. The individuals of the first part are obtained from the non-dominated solutions (from the first front to the fifth front solution) discovered by the GA which is considered as an elitist method (X). We use a parameter 'copy rate from pareto' to represent percentage of population copied from non-dominated solutions. The individuals from the second part are created randomly. It is described by another parameter 'rate of random selection' (Z). The

individuals of the third part come from the previous population via selection, crossover and mutation (Y). The following is an example to demonstrate how to form a population. Supposed the population size is 100, the ‘copy rate from pareto’ is 0.2 and the ‘random created rate’ is 0.1, which means 20 (100*0.2) individuals are copied from the non-dominated solutions, 10 (100*0.1) individuals are created randomly, and the rest 70 (100-20-10) individuals are created from the previous population.



Figures 5: Formation of population

(2.3) Selection method

We use rank-based roulette wheel selection (Kumar and Jyotishree, 2012) to select individuals as parents to create off springs. The mapping function $g(pos)$ of the ranks of individuals for determining their selection probability is as follows:

$$g(pos) = 2 - 2 * (SP - 1) * \frac{pos-1}{n-1} \quad (1)$$

Where SP is the selective pressure that is considered as one parameter of the GA ($1 < SP \leq 2$), pos is the position of the sorted individual in the population P , n is the number of the individuals in the population. This means the higher the position of an activity, the higher rank the activity has.

(2.4) Crossover method

In this algorithm we employ parameterized uniform crossover (Jaddan et al., 2015) to create offspring. After two individuals are selected as parents, at each gene a biased coin is tossed to determine which parent will transmit the gene to the offspring. It assumes that a toss of head will choose the gene from the first parent, and a toss of tail will select the gene from the second parent. The following table shows the mode selection for a tail rate of 0.3. It means that the probability to toss a tail is 0.1-0.3 while the probability to toss a head is 0.4-1.0.

Table 1: Crossover method

Coin toss	Head (0,5)	Tail (0,1)
Parent A	Mode 1	Mode 3
Parent B	Mode 2	Mode 4
Child	Mode 1	Mode 4

(2.5) Solving genetic drift problem

Due to multi-dimensional search space, multi-objective genetic algorithm has an inherent characteristic called genetic drift. The population tends to form relatively few clusters that prevent diverse population. In order to increase diversification, we have utilized the concept of density estimation by means of calculating the total Euclidean distance (figures 6) of a solution to other solutions in the same Pareto front.

The first step is to calculate the Euclidean distance for every solution pair x and y .

$$z(x, y) = \sqrt{\sum_{k=1}^k \left(\frac{z_k(x) - z_k(y)}{z_k^{max} - z_k^{min}} \right)^2} \quad (2)$$

with

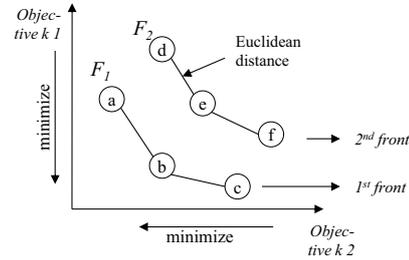
k : amount of objectives

z_k^{min}, z_k^{max} : Maximum and Minimum value of objective function z_k observed so far.

The second step is to calculate the total distance TD of solution x to other solutions t of population P (eq. 3) while the third step is to adjust fitness of solution x (eq. 4):

$$TD(x, t) = \sum_{t \in P} z(x, t) \quad (3)$$

$$f'(x) = \frac{pos}{TD(x, t)} \quad (4)$$



Figures 6: Euclidean distance in Pareto fronts

In this study, the considered optimization objectives are minimizing (1) total cycle time C (sum of cycle time C_j of each product j), (2) total tardiness T (sum of tardiness T_j of each product j referred to the due date d_j), and (3) total costs M (activity-based costing, Cost M_j for manufacturing a product j).

$$C = \min \sum_{j=1}^J C_j \quad (1)$$

$$T = \max \sum_{j=1}^J (C_j - d_j, 0) \quad (2)$$

$$M = \min \sum_{j=1}^J M_j \quad (3)$$

where;

j Product ($j=1 \dots J$)

The unit of tardiness and cycle time is time unit (TU), the unit of costs cost unit (CU).

The applied cost function is a result of the research project of Carl and Angelidis (2015).

COMPUTATIONAL EXPERIMENTS

For answering the research question we conduct computational experiments. The investigated input parameters of the optimizer and the associated DOE is shown in the following table:

Table 2: Full 3-level factorial plan

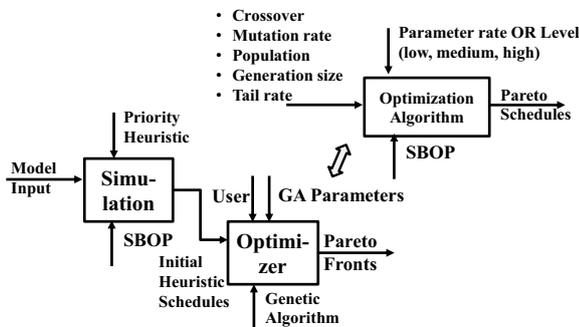
	Low	Medium	High
GS	10	30	50
PS	10	30	50
CR	0.1	0.5	0.9
MR	0.1	0.5	0.9
TR	0.1	0.5	0.9

GS=Generation size, PS=Population Size, CR=Crossover, MR=Mutation rate, TR=Tail rate

Each parameter of the optimizer is investigated with 4 different priority rules, which are:

- First in First out (FIFO)
- Earliest operation Due Date (ODD)
- Minimum Slack (MSLK)
- Shortest processing time (SPT)

So we apply a 3-level plan for the 4 different priority rules which becomes a total of 972 design experiments described by an IDEF diagram in figure 7.



Figures 7: Experiment design

The amount of simulation runs for each experiment depends on the parameters generation size and population size. So one experiment can have in our example up to 2500 simulation models (GS=50, PS=50). So all in all the experiments took 12 days on an 2,49 GHz Intel® Xeon processor with 24 GB RAM.

For each experiment, the best schedule according to the fitness-value $f(x)$ is selected for the investigation.

The following table shows an overview for the results of the best schedules from each experiment:

Table 3: Results of Experiment Runs

	Tardiness (TU)	Cycle Time (TU)	Costs (CU)
Average	2693.37	5456.11	332503.3
Standard deviation	20.34	47.7	818.35
Minimum	2607	5305	322663.9
Maximum	2751	5569	334862.8

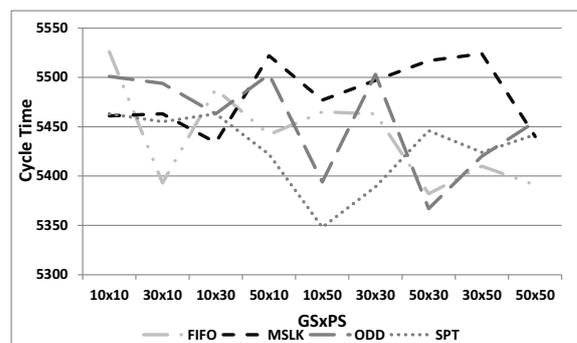
If we have a look on the results, there are significant differences between the minimum and the maximum values for the objectives. To find out which parameter has the strongest influence on the objectives, we have performed a sensitivity analysis.

The following table shows the main effects of the parameters on the single objectives (uncoded parameters). The main effect of a factor is defined as the average change in the output, which is generated by conversion of this factor from a level to the next higher level (for priority rules from random rule to specified rule).

Table 4: Results of Experiment Runs

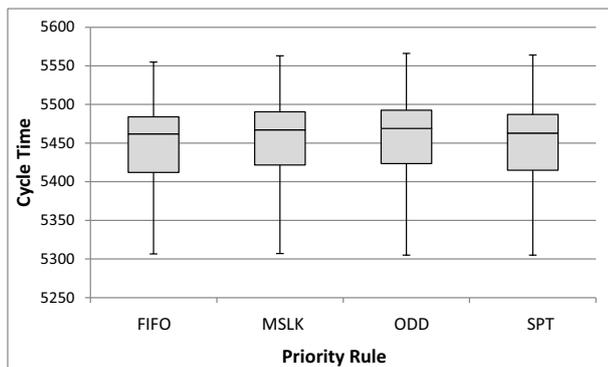
	Tardiness (TU)	Cycle Time (TU)	Costs (CU)
GS	-3.356	-22	-912.8
PS	-15.6	-32.2	-995.98
CR	1.17	17.30	311.26
MR	11.08	4.33	382.92
TR	2.43	6.29	119.28
PR FIFO	-1.1764	-24.18	-65.22
PR MSLK	7.0108	22.5	240.87
PR ODD	1.1842	13.41	142.46
PR SPT	-7.0185	-11.74	-318.12

Table 4 shows that GS and PS are minimizing the objective functions when they are increased while CR, MR, TR are maximizing the objective functions when they are increased. For the usage of the priority rules, the impact on the objective function is different and not so significant. On ranking the impact we observed that in general, GS and PS have maximum influence on objectives although, detail analysis shows that it is not necessary to have continuous improvement by increasing GS and PS. It can be observed in the figure below that with increased GS and PS levels, cycle time does not shows continuous improvement (CR=0,1, MR=0,1 and TR=0,1):



Figures 8: Cycle time to priority rules and GSxPS

The figure shows that it's possible to get with any priority rule good results, but not according to a specific parameter combination. The only exception is MSLK, where results are generally poor. With more simulation runs it tends to better results. So the question is how to get with a high probability good results when choosing a priority rule. For this investigation, we must have a look at the distribution of results, shown in the following figure for cycle time as a box-plot whisker diagram.



Figures 9: Cycle time to priority rules

The median of FIFO and SPT is lower than for MSLK and ODD. For FIFO and SPT, the lower quartile has a wider range than for MSLK and ODD, what means that the possibility to get better results is higher when using FIFO and SPT. Another presumption is, that for this priority rules the impact of the other parameters is much lower. So that means, when we are using SPT or FIFO, it is more likely to get high-quality results with fewer simulation runs. This presumption is part of our further research.

CONCLUSIONS

In this paper, we have investigated various parameters for solving and optimizing the MMRCMPSP, especially the problem domain of complex assembly lines. For solving and optimizing, we used a platform based on the theory of simulation based optimization. GA was used for optimization purposes with initial population generated from priority heuristic scheduling. These two strategies are part of the developed and presented optimizer. The investigation is based on a case study. We tried to find out, which genetic operators have the strongest influence on the search direction of the genetic algorithm regarding the objectives tardiness, cycle time and costs. So we used a 3 level plan for the GA-parameters which were observed in combination with the priority rules. It was observed that GS and PS have the biggest influence on the objective function. Priority rules have different impacts. The probability to get the best result is most likely, when PS and GS are maximal and CR, MR and TR is minimal. For this model, we get the more high quality results from FIFO and SPT without considering the parameters of GA. Hence, conclusion that can be drawn is that in case, priority

heuristics provide poor initial population, higher optimization runs have to be performed for the same results obtained from better initial solutions via better priority rules. Our further investigation will concentrate on methods to find the optimal parameter setting, to get a good trade-off between run time and results. We will also investigate the performance of priority rules in case of different production layouts.

ACKNOWLEDGEMENTS

We would like to thank Evangelos Angelidis and Daniel Bohn for providing the Software SBOP and for their implementation work. The presented work is a result of the research project "Simulationsbasierte dynamische Heuristik zur verteilten Optimierung komplexer Mehrziel-Multiprojekt-Multiressourcen-Produktionsprozesse" (Founded by the Deutsche Forschungsgesellschaft (DFG), Duration 01/2013-01/2017).

REFERENCES

- Alvarez-Benitez, J. E.; Everson, R. M. and Fieldsend, J. E. 2005: A MOPSO Algorithm Based Exclusively on Pareto Dominance Concepts. In: David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell et al. (Hg.): Evolutionary Multi-Criterion Optimization, Bd. 3410. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture notes in computer science), S. 459–473.
- Angelidis, E.; Bohn, D.; Rose, O. 2013: A simulation tool for complex assembly lines with multi-skilled resources. Proceedings of the 2013 Winter Simulation Conference, IEEE, pp 2577-2586.
- Boctor, F.F., 1993. Heuristics for scheduling projects with resource restrictions and several resource-duration modes. International Journal of Production Research, 31(11), pp.2547-2558.
- Bowman, E.H., (1959), "The schedule-sequencing problem", Operations Research, Vol 7, pp. 621– 624
- Brucker, P. 2004. "Scheduling Algorithms," Springer-Verlag Berlin Heidelberg, Berlin, Chap. 1.
- Buddhakulsomsiri, J. and Kim, D. S. 2007. Priority rule-based heuristics for multi-mode resource-constrained project scheduling problems with resource vacations and activity splitting. European Journal of Operational Research, 178(2), pp. 374-390.
- Carl, S.; Angelidis, E. 2015: Schlussbericht zu dem IGF-Vorhaben "Simulationsbasierte Prozesskostenrechnung zur Bestimmung kostenminimaler Ablaufalternativen in der Montageplanung bei KMU". Kurztitel: Prozesskostenorientiertes Montageplanungssystem (ProMoPs).
- Cheng, J., Fowler, J. and Kempf, K., 2012. Simulation based multi-mode resource constrained project scheduling for semiconductor equipment installation and qualification. Proceedings of the Winter Simulation Conference (WSC), 9-12 December, Berlin, Germany, IEEE Transactions
- Conway, R.; Maxwell, W.L. and Miller, L.W. 1967. Theory of Scheduling, Dover Publications Inc.
- Deb, K.; Pratap, A.; Agarwal, S. and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm. NSGA-II. In: IEEE Trans. Evol. Computat. 6 (2), S. 182–197. DOI: 10.1109/4235.996017.

- Jaddan Al, O.; Rajamani, L. and Rao, C. R., "Improved Selection Operator for GA", *Journal of Theoretical and Applied Information Technology*, pp 269–277, 2005.
- Kolisch, R. 1995. Project scheduling under resource constrained- Efficient heuristics for several problem classes. *Physica-Verlag Heidelberg, Springer Inc*
- Kühn, M.; Völker, M.; Angelidis, E.; Bohn, D. and Zhou, Z., 2015. Sensitivity analysis of a cost objective function to derive criteria for the simulation based optimization of planning processes. *Simulation and Production in Logistics 2015, Fraunhofer IRB Verlag, Stuttgart, Germany*, pp. 89-99.
- Kumar, Rakesh; Jyotishree 2012: Blending Roulette Wheel Selection & Rank Selection in Genetic Algorithms. In: *IJMLC*, S. 365–370. DOI: 10.7763/IJMLC.2012.V2.146.
- Kuster, J.; Jannach, D. and Friedrich, G., 2007, "Handling alternative activities in resource-constrained project scheduling problems," *Proceedings of the 20th International joint Conference on Artificial Intelligence, San Francisco, USA*, pp. 1960-1965
- Iringova, M.; Vazan, P.; Kotianova, J. and Jurovata, D. 2012. The comparison of selected priority rules in flexible manufacturing system. *Proceedings of the World Congress on Engineering and Computer Science (WCECS)*, 24-28 October, San Francisco, USA.
- Myszkowski, P.B.; Skowronski, M.E. and Podlowski, L. 2013. Novel heuristics solutions for multi-skill resource constrained project scheduling problem. *Proceedings of Federated Conference on Computer Science and Information System (FedCSIS)*, *IEEE Transactions*, pp. 159-166.
- Olsen, R.P. 1971. Can project management be defined? *Project Management Quarterly*, 2(1), 12-14
- Vanhoucke, M. and Maenhout, B. 2005. PSPLib – A nurse scheduling problem library: a tool to evaluate (meta-) heuristic procedures. In: Brailsford, S. and Harper, P., *Operational Research for Health Policy: Making Better Decisions*, *Proceedings for the 31st Annual Meeting of the working group on Operations Research Applied to Health Services*, pp.151-165.
- Zahid, T.; Völker, M. and Schmidt, T. 2015. A practical algorithmic approach towards multi-modal resource constrained multi-project scheduling problems. *Proceedings of the International Conference of Mechanical Engineering and Exposition (IMECE)*, 13-19 November, 2015, ASME, Houston, Texas
- Zahid, T., (2013), "Multi-criteria optimization of process plans for reconfigurable manufacturing systems: An evolutionary approach", Chap: 2, *Masters Thesis, National University of Science & Technology, Pakistan*

AUTHOR BIOGRAPHIES

Mathias Kühn is member of the scientific staff of Prof. Dr.-Ing. habil. Thorsten Schmidt and a Ph.D. student at the chair of logistics engineering, at the department of mechanical engineering, TU Dresden, Germany. He studied mechanical engineering with the specialisation production technology and factory planning. His research interest includes modelling complex production processes and simulation based optimization for complex assemble lines. His email address is: mathias.kuehn@tu-dresden.de.

Taiba Zahid is working since 2013 as a research associate in the working group of facility planning at Technische Universität in Dresden. She is recently working in a group which focuses on providing practical solutions for industries concerning production management, logistics and supply chain management. Her main research aim is to find robust schedules; insensitive to disruptions and can tolerate uncertainties by remaining close to their optimal solutions. Her email address is: taiba.zahid@tu-dresden.de

Michael Völker, born in 1956, studied mechanical engineering at Technische Universität Dresden. He received his doctorate in 1988. His doctoral thesis analyses the use of industrial robots for automated machine charging. In the context of various industrial projects he gained experience as a senior project manager in the planning and commissioning of more than 20 factories in different countries. In addition, he gave guest lectures at various universities. He is currently working as head of the factory planning department at Technische Universität Dresden, Faculty of Mechanical Engineering, Chair of Logistics Engineering. His expertise in teaching and research lies in particular in the planning and design of production systems and factories. Core issues are Digital Factory concepts and the organization and optimization of production processes. His email address is: michael.voelker@tu-dresden.de.

Zhugen Zhou is a member of the scientific staff of Prof. Dr. Oliver Rose at the Chair of Modeling and Simulation, at the Department of Computer Science, University of the Federal Armed Forces Munich, Germany. He received his M.S. degree in computational engineering from Dresden University of Technology and Ph.D. degree in computer science from University of the Federal Armed Forces Munich. His research interests include dispatching concepts for complex production facilities and simulation based optimization for complex assemble lines. His email address is: zhugen.zhou@unibw.de.

Oliver Rose is the professor for Modeling and Simulation at the Department of Computer Science, University of the Federal Armed Forces Munich, Germany. He received his M.S. degree in applied mathematics and Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member INFORMS Simulation Society, ASIM, and GI. His email address is: oliver.rose@unibw.de.

JOB SHOP SCHEDULING WITH FLEXIBLE ENERGY PRICES

Maximilian Selmair
Thorsten Claus
Marco Trost
Department of Business Science
Dresden Technical University
01062 Dresden, Germany
maximilian.selmair@mailbox.tu-dresden.de

Andreas Bley
Department of Mathematics
Kassel University
34132 Kassel, Germany
andreas.bley@uni-kassel.de

Frank Herrmann
Innovation and Competence Centre
for Production Logistics and
Factory Planning (IPF)
OTH Regensburg
93025 Regensburg, Germany
frank.herrmann@oth-regensburg.de



Keywords—*Job Shop Scheduling; Flexible Energy Prices; Energy Efficient Production Planning; Energy Consumption; Standby*

Abstract—The rising energy prices – particularly over the last decade – pose a new challenge for the manufacturing industry. Reactions to climate change, such as the advancement of renewable energies, raise the expectation of further price increases and variations. Regarding the manufacturing industry, production planning and controlling can have a significant influence on the in-plant energy consumption. In this paper, we develop a scheduling method as a linear optimization model with the objective to minimize energy costs in a job shop production system.

INTRODUCTION

Since the industrial revolution, the worldwide economic prosperity depends on the reliable provision of electric energy. Yet the generation of this energy by means of fossil fuels is, as measured by the associated CO₂-emissions, the main contributor to climate change (Finkbeiner et al. 2010). According to the Federal Association for Energy and Water Management, the electricity costs for private customers rose by 85% between the years 2000 and 2010. Within the same period, an increase of 130% was noted for the industrial sector (Bauernhansl et al. 2013). One of the driving factors in this distinct rise are increases in taxes and other charges, such as the EEG reallocation charge (EEG = Erneuerbare-Energien-Gesetz; Renewable Energies Act of Germany). The most of the remunerated electricity under the EEG is traded at spot-markets like the European Energy Exchange (EEX) or the European Power Exchange (EPEX). As supply and demand determine the price, energy tariffs are highly variable over the day. In line with this, methodologies for price predictions for competitive energy markets have been published by Lei and Feng 2012 and others. The spot-markets are trading electricity for the following day (Day-Ahead). Figure 1 shows exemplary the hourly electricity price for the following day - in this case for the 21st of January 2016, with a standard deviation of 20.75 (39.8%). The hourly electricity prices are used in this research to minimize the energy consumption costs by means of intelligent scheduling.

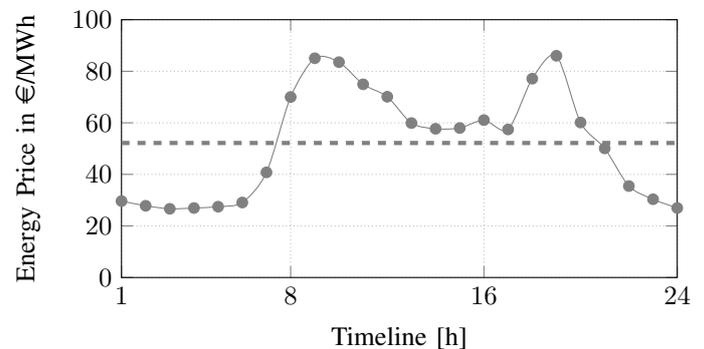


Fig. 1. Hourly electricity price and average (dashed line, for information purposes) for the following day, in this case 21th of January 2016 (Own representation of data from www.epexspot.com)

RELATED LITERATURE

Energy-efficient scheduling and the reduction of energy consumption has been a very important issue over the recent years. In this area of research, Weinert et al. 2011 introduced a so-called energy blocks methodology, which allows for the accurate prediction of energy consumption and integrates energy efficiency criteria into production system planning and scheduling. Dai et al. 2013 proposing an improved genetic simulated annealing algorithm for energy efficient flexible flow shop scheduling, focusing on the two objectives makespan and energy consumption. Furthermore, Liu et al. 2014 developed a multi-objective scheduling method in which the reduction of the energy consumption was one of the primary objectives.

The three papers mentioned above consider only two operational machine states with respect to the energy consumption: Idle (or standby) and processing. In 2014, Shrouf et al. 2014 extended these works by making also decisions on a machine level, which allowed them to consider more operational-modes of a machine. Developing a model for optimizing the total energy consumption costs when scheduling jobs on a single machine, they consider the operating states *Idle*, *Processing*, *Turning Up* and *Turning Down*.

The extension of this approach to more than one machine

complicates matters substantially. Dependencies between all machines are unavoidable and need to be modeled when assuming a job shop production system. Already the basic job shop problem is known to be NP-complete and to be computationally extremely difficult.

Concerning exact solution methods for job shop problems, rather few methods have been published. Until 2005, the most effective approaches have been branch-and-bound algorithms that branch on the job orders on the machines in the so-called disjunctive graph model. In the traditional job shop problem, the optimal starting times of the jobs can be easily computed once the decisions concerning the order of the jobs are made. Aiming to avoid unnecessary branchings, these algorithms typically also employ constraint programming techniques in order to tighten the bounds for the job starting times and infer job orders during the branch-and-bound process.

Motivated by the success of time-indexed models and solution approaches for other scheduling problems (Sousa and Wolsey 1992; Akker 1994), Martin and Shmoys 2005 eventually proposed to use time-indexed integer programming formulations also for the job shop problem. Using such a formulation together with effective bound tightening techniques and specialized branching, they have been able to computationally derive lower bounds that were stronger than those obtained with disjunctive graph models and job order based formulations.

In a time-indexed formulation, the planning horizon is discretized and binary variables are used to indicate if a job starts at a specific time. Formulations of this type are widely used to tackle project scheduling and dynamic planning problems that involve complex resource, precedence, or state constraints, as these additional constraints often can be formulated much easier in a time-index model than in a continuous time model. Already Ford and Fulkerson 1962 observed that dynamic flow problems in a network with transit times on the arcs can be modeled equivalently as static flow problems in time-expanded networks, which is equivalent to a time-indexed formulation of the problem.

Successful applications of time-indexed and time-expanded problem formulations include the optimization of supply chains (Küçükyavuz 2011; Pochet and Wolsey 2006), production planning in mining, energy production, and other industries (Louis and Hill 2003; Chicoisne et al. 2012; Epstein et al. 2012; Lambert et al. 2014), timetabling in transportation (Schöbel 2007; Serafini and Ukovich 1989), and many more. In many of these cases, the time-indexed integer programming formulations also lead to mathematically stronger linear relaxation than their continuous time counterparts, which is beneficial in branch-and-bound algorithms. This benefit typically comes at the cost of a much larger problem formulation. However, exploiting the special structure of the time-indexed formulations in specialized solution algorithms, the size of the formulation that actually has to be solved often can be reduced substantially. A discussion of the main features, strengths, and limitations of alternative modeling and optimization techniques, with a special focus on short-term scheduling of chemical batch processing, can be found in the survey of Méndez et al. 2006.

A computational evaluation of different mixed-integer pro-

gramming formulations for parallel machine scheduling problems for job-related objective functions such as weighted completion time, weighted tardiness, maximum lateness, and number of tardy jobs has been published in Unlu and Mason 2010. The results of this study, as mentioned also in Berghman et al. 2014, suggest that time-indexed formulations perform reliably well for such problems and should be explored further for the solution of scheduling problems with multiple machines. Time-indexed formulations are widely used to model variable operational-modes of devices and plants in various applications (for example in unit commitment planning for electricity networks or in dynamic spectrum assignment in telecommunication networks) or to model time-dependent job-related objective functions in scheduling problems. To the best of our knowledge, however, the use of time-indexed formulations to model the job-independent ramping and switching dynamics of the machines' operational states in a multi-machine scheduling problem has not yet been investigated, yet.

PROBLEM DEFINITION

When considering a common job shop production system, each machine usually has a varying energy demand depending on its operational state. Production systems that consist of chipping (e.g. milling machines) or transforming tool machines (e.g. presses or benders) typically have a vast demand of energy (Neugebauer 2008). Further examples of high energy consumers are industrial laser welding or laser cutting systems (Ahn et al. 2016). Note that a considerable share of the electricity consumption of these machines in practice is actually associated with the standby-mode, when the machines are active but not working (Neugebauer 2008; Ahn et al. 2016). Furthermore, peripheral systems, such as cooling and ventilation, loading and unloading mechanisms, or hydraulic systems require a significant amount of electricity even in standby-mode. Shutting down these modules is generally refused in industrial practice on account of the necessary process stability. Operational states would have to be predictable and reliable in order to initiate a safe ramp down without risking process stability.

If one did assume that machines ramp down entirely when not in use, an initial evaluation would exhibit short idle times and, thus, a high level of machine capacity utilization, which in turn saves energy. This would reduce the energy demand during standby-mode and the machine in question could ramp down after each processing operation. However, long idle times are also possible, which would allow for a complete ramp down of the machine. The feasibility of this option depends on planning a timely and safe restart and the subsequent flawless resumption of production.

Our research specifically addresses these questions. We aim to develop models where the operating-modes of all machines are planned together with the scheduling of the jobs in a period-specific manner such that longer ramp up, ramp down, and standby-processes are adequately considered. Thus, periods with lower energy costs could be utilized to schedule production processes with high energy demands and remaining in standby-mode or even ramping down production facilities in more expensive periods can save energy costs.

Referring to the above mentioned use case (chipping or transforming tool machines as well as laser welding and

cutting), we have identified five crucial operational states that should be considered: *off*, *ramp up*, *setup*, *processing*, *standby* and *ramp down*. ramp up and ramp down can be seen as transitional states with a fixed duration depending on the machine. The transition time between standby and processing or standby and setup and vice versa is assumed to be negligible. In industrial practice, this transition only lasts a matter of seconds and is typically too short to affect a solution that ranges from minutes to hours. The essential decisions related to the machines are to decide whether a machine is switched off and on or whether it is left in standby in a production break. Both choices require energy and cause costs, and the first one is only possible if the break is long enough for ramping down and up.

To determine the processing periods for all operations and the operational states for each machine, our proposed model provides:

- 1) start period of processing each operation on the machines,
- 2) start period for setting up a machine for the upcoming operation (implicitly), and
- 3) all operational status transitions for each machine.

FORMULATION OF THE MODEL

All jobs and machine states are planned within a specific time period. The planning horizon is discretized into $T \in \mathbb{N}$ equally long intervals, called periods, and denoted by $[T] := \{0, \dots, T-1\}$. If ℓ represents the duration of a period, $t \in [T]$ denotes the period from time $t\ell$ to time $(t+1)\ell$. In accordance with Shrouf et al. 2014, every time period is associated with its individual energy price described by $C_t \in \mathbb{R}^+$. Note that all durations and times are given and modeled as integers, so only integer multiples of the period length ℓ can be represented exactly in this model.

The given set of v machines is denoted by $M = \{M_j\}_{j=1}^v$ (using an arbitrary predefined order on the machines). The considered operational machine states are described by the set $S := \{\textit{off}, \textit{standby}, \textit{processing}, \textit{setup}, \textit{rampup}, \textit{rampdown}\}$. For each operational state $s \in S$ and each machine $j \in M$, a specific energy demand $P_{j,s} \in \mathbb{R}$ is given. For the two transition states ramp up and ramp down, we are also given the transition times $d_j^{\textit{rampup}} \in \mathbb{N}$ and $d_j^{\textit{rampdown}} \in \mathbb{N}$ for ramping up machine j from operational state off and for ramping it down to off, respectively.

In accordance with Özgüven et al. 2010, we let $J = \{J_i\}_{i=1}^n$ denote the given set of n jobs.

Each job $i \in J$ consists of $O_i \in \mathbb{N}$ individual operations (sub-tasks). The k -th operation of job i is denoted operation (i, k) . The overall set of all operations of all jobs is denoted by $O := \{(i, k) \mid i \in J, k \in \{1, \dots, O_i\}\}$. For each operation $(i, k) \in O$ we are given

- the machine setup time $d_{i,k}^{\textit{setup}} \in \mathbb{N}_0$,
- the operation processing time $d_{i,k}^{\textit{op}} \in \mathbb{N}$, and
- the associated machine $m_{i,k} \in M$.

Furthermore, for each job $i \in J$ we have

- a release time a_i

- a due time f_i

Note: Release date a_i means job i can start from period a_i (at time $a_i\ell$). Due date f_i means job i must be completed within period $f_i - 1$ (not later than $f_i\ell$).

Assumptions

- 1) Every machine can only process or setup for one operation at a time.
- 2) Once an operation has started to process, interruptions are not allowed. The same applies for setup processes.
- 3) Every job contains operations in a linear sequence. Consequential operation (i, k) must be completed before operation $(i, k+1)$ begins.
- 4) No time is required for changes between operating-modes from standby to processing and vice versa.
- 5) Changes between operating-modes (ramp up and ramp down) cannot be interrupted after they have been initiated.
- 6) A machine can be setup for operation (i, k) even if the preceding operation of the same job $(i, k-1)$ is still being processed on another machine.
- 7) The setup of operations $(i, 1)$ can be initiated prior to the release time a_i of job i .
- 8) Processing operations have to start immediately after the related setup process.
- 9) Two artificial periods are added at the beginning and at the end of the planning horizon (-1 and T), which are free of any machine activity (processing, setup, ramp up or ramp down). These only serve to describe the initial and final states of the machines. In this paper, we assume that all machine must be in state off in these periods.

Preprocessing

Initially, bounds $a_{i,k}$ and $f_{i,k}$ for the earliest and the latest starting times for the individual operations (i, k) , respectively, are determined on the basis of the given parameters. This approach reduces the solution space significantly and increases the speed and efficiency of the model.

- 1) For all operations $(i, k) \in O$ determine:

$$a_{i,k} := \max \left(a_i + \sum_{q=1}^{k-1} d_{i,q}^{\textit{op}}, d_{m_{i,k}}^{\textit{rampup}} + d_{i,k}^{\textit{setup}} \right)$$

$$f_{i,k} := f_i - 1 - \sum_{q=k}^{O_i} d_{i,q}^{\textit{op}}$$

- 2) Determine $A := \{(i, k, t) \in O \times [T] \mid a_{i,k} \leq t \leq f_{i,k}\}$ of possible operations-startperiod-pairs. Thus, operation (i, k) can only start between the periods $a_{i,k}, \dots, f_{i,k}$.

Decision Variables

We introduce two types of binary decision variables: α -variables model the start periods of the operations and β -variables represents the operational states for all machines in all periods.

For each operation (i, k) and each start-period t with $(i, k, t) \in A$ (i.e., t is a permissible start time for (i, k)), we

have a binary variable $\alpha_{i,k,t} \in \{0, 1\}$, which is interpreted as

$$\alpha_{i,k,t} = \begin{cases} 1 & \text{Processing of operation } (i, k) \\ & \text{starts in period } t. \\ 0 & \text{Else.} \end{cases}$$

For each machine $j \in M$, each state $s \in S$, and each period $t \in [T] \cup \{-1, T\}$, we have a binary variable $\beta_{j,s,t} \in \{0, 1\}$, which means

$$\beta_{j,s,t} = \begin{cases} 1 & \text{In period } t \text{ machines } j \\ & \text{is in operational state } s. \\ 0 & \text{Else.} \end{cases}$$

Objective Function

The objective function needs to determine and minimize the energy consumption costs. The operational state of each machine is set by the decision variable β . Parameter $P_{j,s}$ represents the associated power demand. With C_t the energy price per period is provided. Thus Equation 1 minimizes the total energy costs.

$$\min \left(Z = \sum_{j \in M} \sum_{t=0}^{T-1} \sum_{s \in S} \beta_{j,s,t} \cdot P_{j,s} \cdot C_t \right) \quad (1)$$

Constraints

Equalities (2) ensure that every machine has exactly one operational state in each period.

Equalities (3) fix the specific operational state off at the beginning (period -1) and in the end (period T) of the planning horizon for each machine.

Equalities (4) ensure that every operation will start exactly once in its permissible horizon (depending on the release and due date).

Inequalities (5) ensure that machine j is in operational state processing in period t if some operation of duration d started between $t-d+1$ and t and, thus, is still running in period t on this machine. Similarly, inequalities (6) ensure that machine j is in operational state setup in period t if some operation with setup time d starts between $t+1$ and $t+d$ and, thus, requires machine setup in period t on this machine. Moreover, together with (2) these constraints guarantee that machine j can be in setup-mode for or actually executing at most one single operation at a time. Thus, operations and setups do not overlap on any machine, the so-called parallel constraints hold.

Inequalities (7) imply the so-called sequential constraints. Enforcing for all times t that operation (i, k) starts no later than $t - d_{i,k}^{processing}$ if operation $(i, k+1)$ starts in period t (or earlier), these inequalities imply that operation (i, k) indeed completes running before operation $(i, k+1)$ starts.

Inequalities (8) and (9) finally model the technical constraints that are related to the machine states and the duration of ramp up and ramp down phases. The required minimum duration of the ramp down phases is enforced via constraints (8). These ensures that, if machine j is active (i.e. processing, in setup, or in standby) in period t , then it cannot be off (or even already in ramp up-mode again) in period $t + d_j^{rampdown}$

(or earlier): It must either remain active in processing, setup, or standby-mode after the operation it was executing (or setting up for) in period t or, if it decides to ramp down after this operation, the ramp down phase cannot have ended by period $t + d_j^{rampdown}$ or earlier. Similarly, constraints (9) ensure that the ramp up phases are at least as long as required. If the energy consumption in the ramp up and ramp down states is not lower than that in the off state and, similarly, that energy consumption in the processing and setup state is not lower than that in the standby state, these constraints suffice to ensure that the machine state schedules in an optimal solution of the model satisfy the given constraints. Otherwise, one may add further constraints similar to (8) and (9) to ensure that ramping phases have exactly the required lengths and that machines actually switch to off or standby whenever possible.

$$\sum_{s \in S} \beta_{j,s,t} = 1 \quad (2)$$

$$\forall j \in M, t \in [T] \cup \{-1, T\}$$

$$\beta_{m,off,t} = 1 \quad (3)$$

$$\forall m \in M, t \in \{-1, T\}$$

$$\sum_{t \in [T]: (i,k,t) \in A} \alpha_{i,k,t} = 1 \quad (4)$$

$$\forall (i, k) \in O$$

$$\sum_{\substack{(i,k) \in O: \\ m_{i,k}=j}} \sum_{q=t-d_{i,k}^{processing}+1}^t \alpha_{i,k,q} \leq \beta_{j,processing,t} \quad (5)$$

$$\forall j \in M, t \in [T]$$

$$\sum_{\substack{(i,k) \in O: \\ m_{i,k}=j}} \sum_{q=t+1}^{t+d_{i,k}^{setup}} \alpha_{i,k,q} \leq \beta_{j,setup,t} \quad (6)$$

$$\forall j \in M, t \in [T]$$

$$\sum_{q=0}^{t-d_{i,k}^{processing}} \alpha_{i,k,q} \geq \sum_{q=0}^t \alpha_{i,k+1,q} \quad (7)$$

$$\forall i \in J, k \in \{1, \dots, O_i - 1\}, t \in [T]$$

$$\beta_{j,off,q} + \beta_{j,rampup,q} \leq \quad (8)$$

$$1 - \beta_{j,processing,t} - \beta_{j,setup,t} - \beta_{j,standby,t}$$

$$\forall j \in M, t \in [T], q \in \{t+1, \dots, t+d_j^{rampdown}\}$$

$$\beta_{j,off,q} + \beta_{j,rampdown,q} \leq \quad (9)$$

$$1 - \beta_{j,processing,t} - \beta_{j,setup,t} - \beta_{j,standby,t}$$

$$\forall j \in M, t \in [T], q \in \{t-d_j^{rampup}, \dots, t-1\}$$

COMPUTATIONAL RESULTS

This section presents an exemplary case study of a 5×5 job shop problem to demonstrate how scheduling affects the total energy consumption and total energy costs. The study scrutinizes five jobs processed on the same number of machines. The planning horizon spans three consecutive days. It was decided to plan by hours and every period lasts one hour with a total of 72 periods. The proposed plans rely on the energy price model given in Figure 1 for each day. Consequential energy is most expensive between 8 a.m. and 8 p.m.. Our proposed planning horizon begins and ends at midnight. All jobs and their respective release and due dates are given in Table I. These dates are to be strictly adhered to, as delayed jobs are not allowed. The associated operations with all related parameters are given in Table II.

i	a_i	f_i
1	0	72
2	8	72
3	16	72
4	24	72
5	48	72

(i, k)	$m_{i,k}$	$d_{i,k}^{setup}$	$d_{i,k}^{processing}$
1, 1	1	3	4
1, 2	2	3	4
1, 3	4	1	6
1, 4	5	1	6
1, 5	2	4	4
2, 1	3	3	4
2, 2	2	3	4
2, 3	5	1	5
2, 4	4	1	5
2, 5	1	3	4
3, 1	1	4	5
3, 2	2	4	5
3, 3	3	4	8
3, 4	5	3	4
4, 1	3	2	5
4, 2	2	2	5
4, 3	4	1	4
4, 4	5	1	4
5, 1	1	2	3
5, 2	2	2	3
5, 3	3	2	3

j	1	2	3	4	5
d_j^{rampup}	3	3	3	2	1
$d_j^{rampdown}$	2	2	2	1	1
$P_{j,off}$	0	0	0	0	0
$P_{j,rampup}$	18	10	5	4	2
$P_{j,setup}$	8	8	8	3	3
$P_{j,processing}$	20	20	20	6	6
$P_{j,standby}$	7	1	0.5	0.5	0.5
$P_{j,rampdown}$	5	5	5	2	2

As presented by Table III, the duration for ramping up and down as well as the demand for energy in the different operational states varies between machines. Machine M_1 , for example, has the highest energy consumption in standby-mode. Ramping up is also quite expensive in comparison to the other machines of the production system. Machines M_3 – M_5 require less energy and are comparatively cheap in standby-mode. The highest consumption of energy for processing and setup-mode is linked with Machines M_1 – M_3 . It is expected that our model will schedule jobs to these machines only in periods with cheap energy prices, if possible.

Figure 3 visualizes a schedule plan without taking either energy consumption or energy prices into consideration. All

jobs are planned by minimizing their makespan to complete them as soon as possible. Along with the planned operational periods, all further machine-specific operating-modes are visualized. The key can be found in Figure 5.

Figure 4 presents the energy-efficient solution of our new model. Several things are particularly noticeable. The first salient findings are the scheduled operational states. The machines are not switched on continuously. In addition to the setup and processing states, ramping up and down is planned as well as the standby-mode. The analysis of the schedule of M_1 – M_3 was the first step. As shown in Table III, these machines have a vast demand for energy in all operational states. M_1 has the highest energy consumption in standby-mode. This is reflected by the schedule plan: M_2 and M_3 ramp up hours before they start to process operations. This can be explained by the energy prices. As energy is cheap between 0 a.m. and 8 a.m., the model plans expensive processes in such periods. Obviously the cost for the subsequent standby-mode over many hours is lower than ramping up the machines just prior to the job. This was also observed for the ramping down of M_3 . M_1 ramps up just in time due to its high energy consumption during standby-mode. Consequently the standby-mode for M_1 is used very rarely. M_3 is in standby-mode during the more expensive periods. In contrast, M_1 and M_2 are processing during these expensive periods as specific due dates need to be met. M_5 does not use the standby-mode. Although energy consumption in standby-mode is very low, it is cheaper to turn the machine off completely during the non-productive time.

The key performance indicators for both solutions are compared in Figure 2. It is interesting to note that, with exception of M_1 , the energy consumption of the optimized solution remains the same or is indeed higher than its makespan counterpart. Yet the resulting energy costs are lower owing to the well-conceived scheduling strategy. Merely M_4 causes slightly higher costs in our model compared to the minimizing makespan model.

Table IV aggregates the energy consumption and the resulting costs for all machines of scenario 1 (optimized makespan) and scenario 2 (optimized energy consumption costs). The provided significant savings are given in the last two columns.

	Scenario 1	Scenario 2	Savings	
energy consumption	2,194 kWh	2,052 kWh	142 kWh	6.5%
energy costs	€120	€93	€27	22.3%

CONCLUSIONS AND FUTURE WORK

This work proposed a model for minimizing the total energy consumption costs when scheduling a job shop production system. Considering the continuous changes of energy prices, our model can help to organize a more efficient production schedule, especially for high-energy production systems. Furthermore we evaluated the significant energy price savings that could be obtained by using this model instead of the commonly used lead time minimization.

For further benchmark experiments, we propose to use the model for a continuous rolling and overlapping planning long-term study by means of simulation. Finally, our study

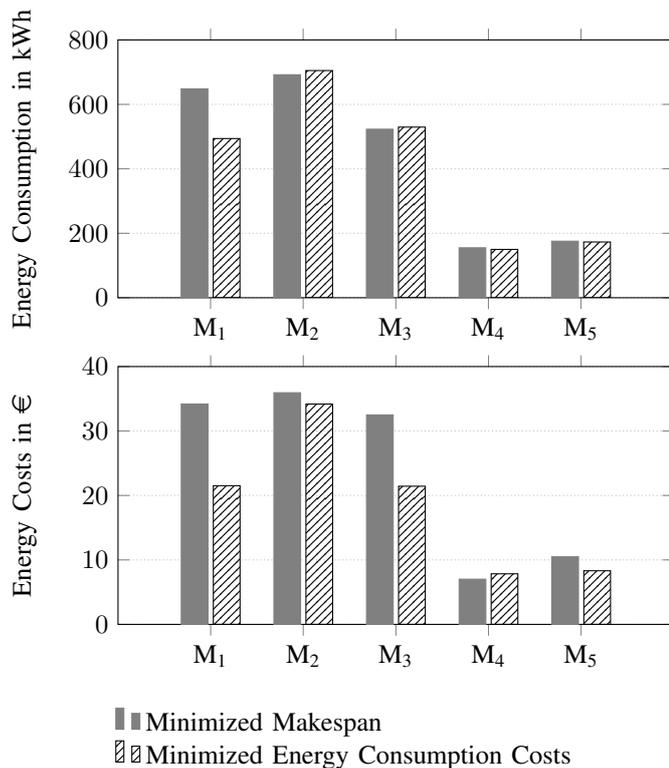


Fig. 2. Comparison of Schedule Plans in Terms of Energy Consumption and Costs

is planned to be integrated as an ecological component of a sustainable production planning concept. The hierarchical production planning as proposed by Hax and Meal 1973 might contribute to creating an ecological and also social environment for sustainable production planning (Trost et al. 2016).

REFERENCES

Ahn, Jong Wook, Wan Sik Woo, and Choon Man Lee (2016). "A study on the energy efficiency of specific cutting energy in laser-assisted machining". In: *Applied Thermal Engineering* 94, pp. 748–753.

Akker, J. M. (1994). *LP-based solution methods for single-machine scheduling problems*.

Bauernhansl, Thomas, Jörg Mandel, Sylvia Wahren, Robert Kasprowicz, and Robert Mieke (2013). *Energieeffizienz in Deutschland: Ausgewählte Ergebnisse einer Analyse von mehr als 250 Veröffentlichungen*.

Berghman, Lotte, Frits Spieksma, and Vincent T'Kindt (2014). "Solving a Time-Indexed Formulation by Preprocessing and Cutting Planes". In: *SSRN Electronic Journal*.

Chicoisne, R., D. Espinoza, M. Goycoolea, E. Moreno, and E. Rubio (2012). "A new algorithm for the open-pit mine scheduling problem." in: *Operations Research* 60, pp. 517–528.

Dai, Min, Dunbing Tang, Adriana Giret, Miguel A. Salido, and W. D. Li (2013). "Energy-efficient scheduling for a flexible flow shop using an improved genetic-simulated annealing algorithm". In: *Robotics and Computer-Integrated Manufacturing* 29.5, pp. 418–429.

Epstein, Rafael, Marcel Goic, Andrés Weintraub, Jaime Catalán, Pablo Santibáñez, Rodolfo Urrutia, Raúl Cancino, Sergio Gaete, Augusto Aguayo, and Felipe Caro (2012). "Optimizing Long-Term Production Plans in Underground and Open-Pit Copper Mines". In: *Operations Research* 60.1, pp. 4–17.

Finkbeiner, Matthias, Erwin M. Schau, Annekatriin Lehmann, and Marzia Traverso (2010). "Towards Life Cycle Sustainability Assessment". In: *Sustainability* 2.10, p. 3309.

Ford, L. R. and D. R. Fulkerson (1962). *Flows in networks*. Princeton landmarks in mathematics. Princeton, N.J. and Woodstock: Princeton University Press.

Hax, Arnaldo C. and Harlan C. Meal (1973). *Hierarchical integration of production planning and scheduling*.

Küçükyavuz, Simge (2011). "Mixed-Integer Optimization Approaches for Deterministic and Stochastic Inventory Management: 7". In: *Tutorials in Operations Research*. Ed. by Joseph Geunes, Paul Gray, and Harvey J. Greenberg. INFORMS, pp. 90–105.

Lambert, Brian W., Andrea Brickey, Alexandra M. Newman, and Kelly Eurek (2014). "Open-Pit Block-Sequencing Formulations: A Tutorial". In: *Interfaces* 44.2, pp. 127–142.

Lei, Mingli and Zuren Feng (2012). "A proposed grey model for short-term electricity price forecasting in competitive power markets". In: *International Journal of Electrical Power & Energy Systems* 43.1, pp. 531–538.

Liu, Ying, Haibo Dong, Niels Lohse, Sanja Petrovic, and Nabil Gindy (2014). "An investigation into minimising total energy consumption and total weighted tardiness in job shops". In: *Journal of Cleaner Production* 65.1, pp. 87–96.

Louis, Caccetta and Stephen P. Hill (2003). "An Application of Branch and Cut to Open Pit Mine Scheduling". In: *Journal of Global Optimization* 27.2, pp. 349–365.

Martin, Paul and David B. Shmoys (2005). "A new approach to computing optimal schedules for the job-shop scheduling problem". In: *Integer Programming and Combinatorial Optimization*. Ed. by WilliamH. Cunningham, S.Thomas McCormick, and Maurice Queyranne. Vol. 1084. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 389–403.

Méndez, Carlos A., Jaime Cerdá, Ignacio E. Grossmann, Iiro Harjunkoski, and Marco Fahl (2006). "State-of-the-art review of optimization methods for short-term scheduling of batch processes". In: *Computers & Chemical Engineering* 30.6–7, pp. 913–946.

Neugebauer, Reimund (2008). *Untersuchung zur Energieeffizienz in der Produktion*. Fraunhofer Gesellschaft.

Özgüven, Cemal, Lale Özbakır, and Yasemin Yavuz (2010). "Mathematical models for job-shop scheduling problems with routing and process plan flexibility". In: *Applied Mathematical Modelling* 6.34, pp. 1539–1548.

Pochet, Yves and Laurence A. Wolsey (2006). *Production planning by mixed integer programming*. Springer series in operations research and financial engineering. New York and Berlin: Springer.

Schöbel, Anita (2007). "Integer Programming Approaches for Solving the Delay Management Problem". In: *Algorithmic Methods for Railway Optimization*. Ed. by Frank Geraets, Leo Kroon, Anita Schoebel, Dorothea Wagner, and Christos Zaroliagis. Vol. 4359. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 145–170.

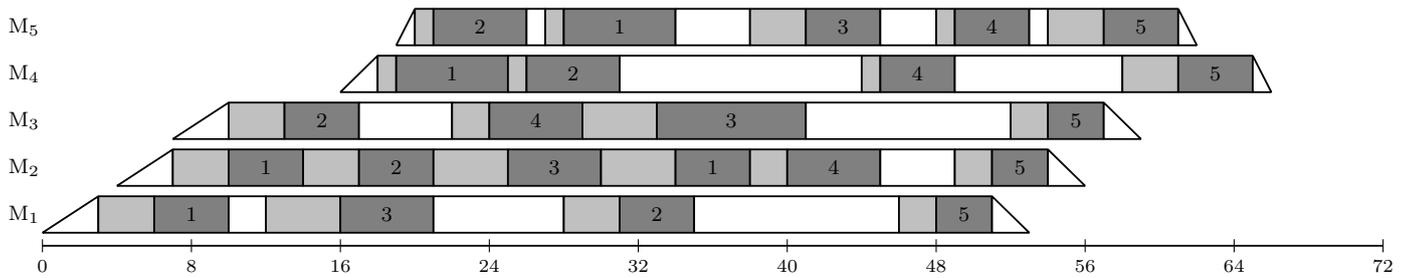


Fig. 3. Schedule plan for minimized makespan

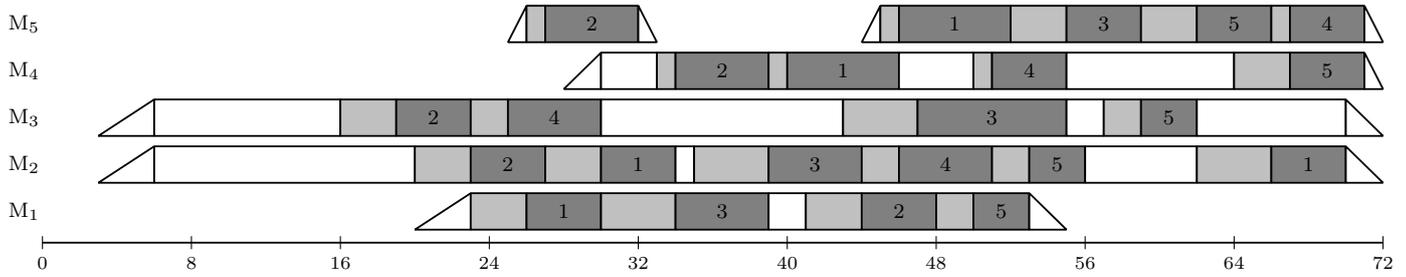


Fig. 4. Schedule plan for minimized energy consumption costs



Fig. 5. Key for Figures 3 and 4

Serafini, Paolo and Walter Ukovich (1989). “A Mathematical Model for Periodic Scheduling Problems”. In: *SIAM Journal on Discrete Mathematics* 2.4, pp. 550–581.

Shrouf, Fadi, Joaquin Ordieres-Meré, Alvaro García-Sánchez, and Miguel Ortega-Mier (2014). “Optimizing the production scheduling of a single machine to minimize total energy consumption costs”. In: *Journal of Cleaner Production* 67, pp. 197–207.

Sousa, Jorge P. and Laurence A. Wolsey (1992). “A time indexed formulation of non-preemptive single machine scheduling problems”. In: *Mathematical Programming* 54.1, pp. 353–367.

Trost, Marco, Thorsten Claus, Frank Herrmann, Enrico Teich, and Maximilian Selmair (2016). “Social and Ecological Capabilities for a Sustainable Hierarchical Production Planning”. In:

Unlu, Yasin and Scott J. Mason (2010). “Evaluation of mixed integer programming formulations for non-preemptive parallel machine scheduling problems”. In: *Computers & Industrial Engineering* 58.4, pp. 785–800.

Weinert, Nils, Stylianos Chiotellis, and Günther Seliger (2011). “Methodology for planning and operating energy-efficient production systems”. In: *5CIRP6 Annals - Manufacturing Technology* 60.1, pp. 41–44.

AUTHOR BIOGRAPHIES

Maximilian Selmair is doctoral student at the Department of Business Science at the Dresden Technical University. Currently employed at the SimPlan AG, he is in charge of projects in the area of material flow simulation. His email address is: maximilian.selmair@mailbox.tu-dresden.de and his website can be found at maximilian.selmair.de.

Prof. Dr. Thorsten Claus holds the professorship for Production and Information Technology at the International Institute (IHI) Zittau, a central academic unit of Dresden Technical University. His e-mail address is: thorsten.claus@tu-dresden.de.

Prof. Dr. Frank Herrmann holds the professorship for information systems in the department of informatics and mathematics at the Regensburg Technical University of Applied Sciences and he is the head of the Innovation and Competence Centre for Production Logistics and Factory Planning (IPF). His e-mail address is: frank.herrmann@oth-regensburg.de.

Prof. Dr. Andreas Bley is professor for applied discrete mathematics at the University of Kassel. His e-mail address is: andreas.bley@uni-kassel.de.

Marco Trost is doctoral student at the Department of Business Science at the Dresden Technical University and he is sponsored by the European Social Fund (ESF). His e-mail address is: marco.trost@mailbox.tu-dresden.de.

MODEL-BASED APPROACH TO STUDY HOT ROLLING MILLS WITH DATA FARMING

Dariusz Król, Renata Słota, Jacek Kitowski
Academic Computer Centre Cyfronet AGH
and Department of Computer Science
AGH University of Science and Technology
ul. Nawojki 11, 30-950 Krakow, Poland
E-mail: dkrol@agh.edu.pl

Łukasz Rauch, Krzysztof Bzowski, Maciej Pietrzyk
Department of Applied Computer Science
and Modelling
and Academic Computer Centre Cyfronet AGH
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Krakow, Poland

KEYWORDS

Model-based simulations, parameter studies, high-performance computing, rolling technology design.

ABSTRACT

The paper describes a computer system for simulating metallurgical rolling processes that consist of multiple steps, each of which is performed by a different type of devices. Both devices and processed materials are described with models, which can be dynamically reconfigured between simulation runs to study different device and environment configurations. Such an approach is especially crucial in technology design based on multi-iterative optimization procedures, for which an objective function uses computationally intensive algorithms. Due to the approach proposed in this paper, in the first stage of optimization more general and coarse models can be applied characterized by lower predictive capabilities and higher computational efficiency. Afterwards, when the optimization procedure finds a solution close to the optimal one, very detailed models can be used to obtain high quality solutions in the last few steps of calculations. To achieve such an objective a hybrid computer system able to use High Performance Computing (HPC) infrastructures was designed and implemented. The details of proposed approach are described, which is followed by presentation of a data farming platform responsible for distribution of complex numerical simulations onto various computer clusters. Finally, a concrete use case of a hot rolling mill is presented and analyzed.

INTRODUCTION

Computer simulation is an essential tool used to verify a stated hypothesis faster and in a more cost-effective way compared with physical experiments in various computational-oriented science fields and in the industry. Metallurgy and metal forming processes, e.g. rolling and cooling, are representative examples of successful simulation applications. Attaining desired mechanical properties of hot rolled Advanced High Strength Steels (AHSS) and Ultra High Strength Steels (UHSS) strips (which are utilized in many branches of industry, e.g. the automobile one) requires implementation of thermo-mechanical schemes, in terms of time/temperature/deformation along the hot rolling process. Such a complex metallurgical

process can be realized in a cost-effective way by simulating hot rolling strip mills consisting of multiple devices with distinct configurations.

Data farming (Horne and Seichter 2014) is an example of a methodology, which combines data exploration and analysis methods with efficient exploitation of modern computational infrastructures such as HPC computer clusters. Its main goal is to increase data volume in a systematic manner in virtual experiments by executing the same simulation many times with different input parameter values. The collected results are then used to gain knowledge about the studied processes.

Advancements in computer hardware in recent years have significantly decreased the time required to run simulations and enabled refinement of simulation models with regard to their complexity (Rauch 2012). Besides accelerating simulations, the modern high-performance computer clusters are capable of processing much more data in a given time interval than ever before. As a result, data farming experiments containing dozens of simulation cases can be finally conducted during the time, which is satisfactory for the industry.

However, harnessing computer clusters to execute complex numerical simulations is a challenging task when dealing with large-scale simulation-based experiments. Neither remote access nor unified interface nor provision of abstraction layers are typically offered by current queuing systems to defining computer experiments as a collection of simulation runs with different input parameter values, like different material models or device configurations. Thus, in the presented work, a higher-level platform, called Scalarm, for data farming computing (Król and Kitowski 2016) has been used.

The above mentioned challenges justified the development of a model-based computer system (VirtRoll) in the framework of the Research Fund for Coal and Steel (RFCS) VirtROLL project, which combines numerical simulations, multiscale modeling, meta-modelling, inverse analysis and optimization techniques to minimize costs of design of production technologies and to optimize semi- and final product properties. In a nutshell, the main objective of the project is to combine a model database and inverse solution coupled with optimization techniques in one comprehensive computer system oriented on enabling domain experts to create and simulate a virtual hot rolling mill (Rauch et al. 2012) equipped with selected

devices.

The rest of the paper is organized as follows: Section 2 specifies the problem, which is solved by the proposed system, Section 3 contains an overview of the proposed system, its architecture and differentiating features, Section 4 describes the system's evaluation in the context of hot rolling mills simulations and Section 5 concludes the paper.

PROBLEM STATEMENT

Let us assume we have a multi-phase metallurgical process to simulate, in which each phase is conducted by a different device. Each device is modelled by a different numerical procedure with multiple configuration options, which influence the results. The processed material is described by constant properties and models, including description of rheology and microstructure evolution. The main problem is to determine the optimal parameters of production devices, which lead to the desired output, i.e. to semi- and final products with specified thermo-mechanical properties. Moreover, the goal is to achieve the answer in reasonable time, which could be accepted also by the industrial practice. Therefore the main objective of the work is to design and implement such mechanisms, which allow for:

- replacing numerical models dynamically between subsequent iterations of an optimization procedure,
- facilitating usage of modern HPC infrastructures in a seamless manner.

PROPOSED SOLUTION

Studying rolling-related processes, by building a virtual hot rolling mill can be described as a multi-step workflow involving: 1) design of a virtual hot rolling mill, 2) design of the computational experiment, 3) simulation of the rolling process with the parameter study approach, and 4) output data exploration with optimization and sensitivity analysis methods to discover relationships between the hot rolling mill parameters and the obtained thermo-mechanical properties in the final product. An overview of the proposed system is depicted in Figure 1. There are three main elements of this solution: a virtual workbench where a hot rolling mill is designed and configured; middleware which is responsible for simulation scheduling onto remote HPC computer clusters; and numerical simulations actually executed on the clusters.

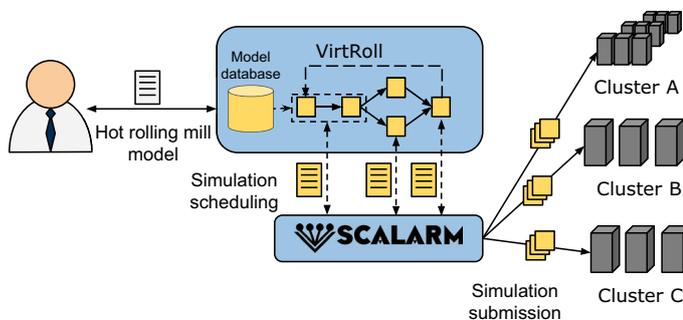


Figure 1: Overview of VirtRoll Integration with the Scalarm Platform.

Design of Hot Rolling Mill – Web-Based VirtRoll Module

The functionality leading to design a new rolling mill is covered by a web-based module, which, in connection with a database, allows for efficient management of production devices, materials and models. Due to the flexibility-oriented design of the database a schema-less document-oriented MongoDB engine was used. This choice was dictated by the necessity to keep the flexible data model of the process supporting addition of new materials and devices characterized by different parameters.

The hot rolling mill design is prepared with a virtual workplace, which enables the users to either prepare a hot rolling mill design from the scratch or select and fine-tune an existing project of a rolling mill common scheme. Moreover, the user can upload new material or devices models and use them in the simulation process. At first, the user selects devices, their placement on the rolling line and configuration parameters. Afterwards, the material to be processed and its models are selected.

The database includes descriptions of different kinds of steels, characterized by grade, name and chemical composition, which can be used further by numerical models.

Flexibility of the VirtRoll system enables the users to add new material models through the graphical user interface (GUI). Therefore, a generic library with abstract classes organized in a multi-level hierarchy for direct inheritance was created. The classes already implemented are as follows:

- **Model** – a root class of the whole hierarchy, containing abstract method for the main algorithmic part of the model. The method is implemented by each model separately. The main attributes of this class are collection of parameters and map of the parameters, which allow to manage model parameters dynamically.
- **RheologicalModel** – the main class for rheological models containing definition of fundamental models for plastic metal forming. Specific models like `HenselSpittel`, `CEMEF` or `Sellars` inherit from this class and define their own parameters as well as a method for numerical calculations,
- **MicrostructureEvolution** – the abstract class allowing definition of derived classes responsible for calculation of grain growth, static `StaticRecrystallizationModel`, dynamic `DynamicRecrystallizationModel` and meta-dynamic `MetadynamicRecrystallizationModel` recrystallization. The simulations of microstructure evolution are managed by `MicrostructureEvolutionLogic`, which aggregates all other models. The logic class can be also inherited and implemented specifically for new material grades.
- **PhaseTransformation** – defines basic functionality for classes responsible for simulation of four main phase transformations i.e. `FerriteTransformationModel`, `PearliteTransformationModel`, `BainiteTransformationModel` and

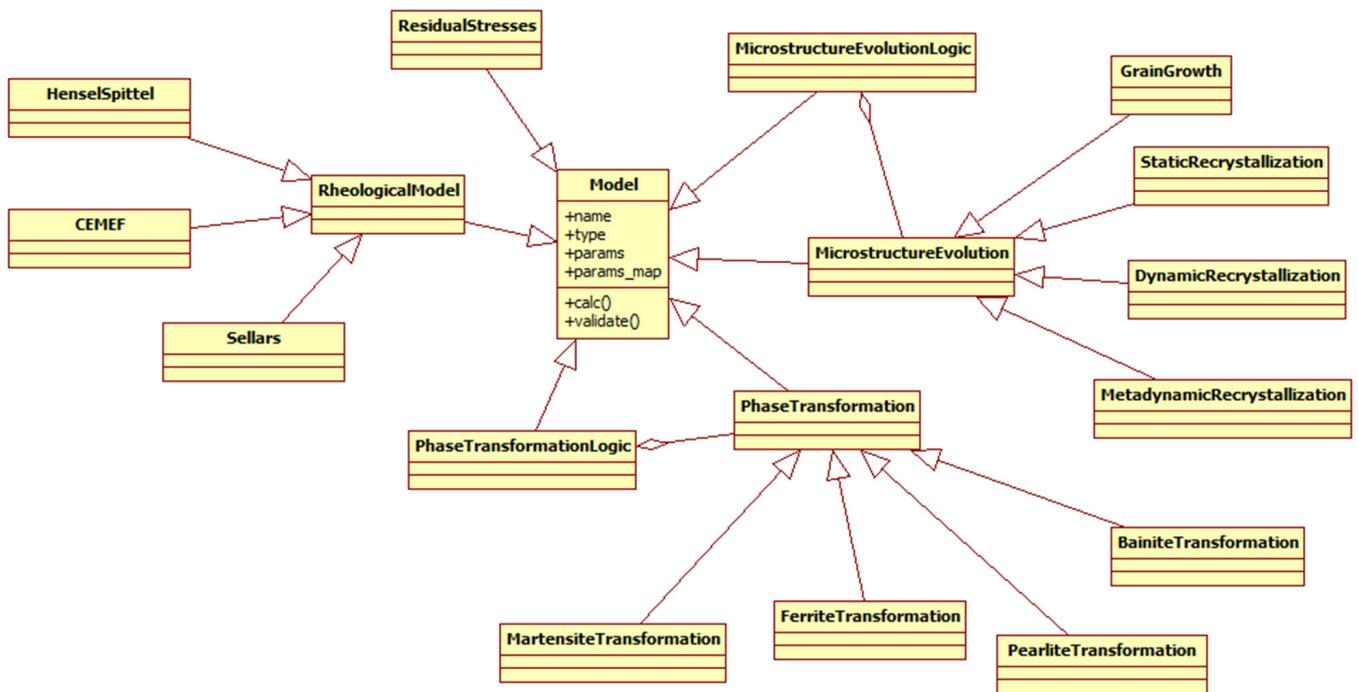


Figure 2: Module Containing Classes Used in Modeling of Hot Rolling.

MartensiteTransformationModel. Similarly to microstructure evolution in the case of phase transformation the separated class, dedicated to management of phase transformation logic, was implemented (PhaseTransformationLogicModel).

The implemented models in the form of different classes of the presented hierarchy were described in (Pietrzyk et al. 2015) in more details.

Design of Computational Experiment with Scalarm Platform

In the presented work, verification and validation of semi- and final products properties obtained with the designed virtual hot rolling mill is conducted by applying the data farming methodology. The user prepares a parameter space involving formed material parameters, devices configuration in the rolling mill and environmental conditions. Each data point in the parameter space describes a distinct simulation case of a virtual hot rolling mill. Therefore each data point can be executed in parallel using distributed HPC computer clusters. Simulation results are available online to visualize progress of any given simulation. This functionality is attained by integrating the VirtRoll system with the Scalarm platform. Scalarm is a tool designed to support the management of data farming experiments, including input parameter space specification, execution, and results collection. Scalarm provides two user interfaces: graphical user interface in form of a web-based application and HTTP-based API. The graphical user interface is oriented towards the users who currently run their simulation codes manually on various infrastructures and would like to facilitate this process. The HTTP-based API is exposed to

support integration of third-party tools with Scalarm using the common JSON data representation format. Hence, the integration of the VirtRoll system with Scalarm is based on this approach as depicted in Figure 1. The following list describes the most important methods of the API for computation delegation from the VirtRoll system to actual computational infrastructure:

- (1) registering a simulation scenario – it is a prerequisite of conducting a parameter study experiment with Scalarm; in the presented case the VirtRoll system registers simulation binaries/codes and description of simulation input parameters for each hot rolling process phase,
- (2) starting a new simulation experiment – arguments of this method include identification of a previously registered simulation scenario and input space specification, i.e. parametrization types and specific attributes for each selected parametrization type, e.g. a model describing material to be processed or configuration options for a simulated device,
- (3) scheduling computations – the user can specify how much computing power and from which cluster should be used to execute simulations,
- (4) getting information about experiments progress – at any given time the user can check how many simulation runs are running and how many of them were completed,
- (5) download experiment results – an integrated third-party tool can download results of the already completed simulation runs, e.g. in form of a CSV file, to enable further analysis in the VirtRoll system,
- (6) extending an experiment – based on the conducted results exploration, it may be necessary to explore some additional parameter space to study some interesting cases, not included previously,
- (7) stopping an experiment – when all simulation runs were

completed, the experiment can be stopped and marked as historical; in such a case the experiment results are stored in the platform and can be explored.

The open-source Scalarm platform supports simulation execution on different types of computing infrastructures including: computer clusters, grids and clouds, by using abstraction of computing tasks, infrastructure facades and infrastructure credentials (Król et al. 2014b). It is available online (<http://www.scalarm.com>), hence the third party tools can be integrated and used in any research. Scalarm was started as a module for data farming within the EDA EUSAS project (Kvassay et al. 2012), where it was used to enhance the training process of security forces through evaluating strategies used during missions. Since then, Scalarm has been used in other scientific disciplines including: computational chemistry (Król et al. 2014a), metallurgy (Rauch et al. 2015), and computer science (Funika et al. 2015).

Numerical Simulations on HPC Computer Clusters

The main part of the system is encapsulated in separated computational module responsible for numerical modelling of macro and micro properties of material. The macro scale simulations of temperatures are realized by using FE method coupled with mechanical and microstructural models.

Mechanical Model.

Originally developed finite element (FE) simulation model (Pietrzyk 2000) was used to calculate strains, stresses, forces, torques and temperatures. Even if a simple stationary FE model with a coarse mesh is used in simulations of metal flow in rolling, the computing time for one pass is about 2-3 minutes. Since a lot of passes have to be simulated to determine one value of the objective function, it is useful to search for alternative models, which can accelerate optimization. Application of the metamodel approach is such an alternative. A metamodel of the process or phenomenon is a certain abstraction created on the basis of the lower level model developed using mathematical techniques. Thus, any approximation of the basic model, which gives reasonably realistic description of the process, can be considered as a metamodel, which allows for significant decrease of the computing time.

Various techniques can be used to build metamodels. Artificial Intelligence (AI) methods, in particular Artificial Neural Networks (ANN), are the most common. When the cost of computations for training data is not so high and large training data sets can be created, application of the ANN is efficient and even very complex relationships can be accurately described by the metamodel. Contrary, when the FE method is used to generate training data, the costs of computations of one set of data are high and other metamodeling techniques should be searched. The surface response method was used in the present work to calculate mechanical parameters including strains, stresses, forces and torques. An additional advantage was made of the fact that the material flow stress is the main factor, which decides about the accuracy of calculation of force parameters and influence of the geometrical parameters is of lesser importance. Therefore, the emphasis was put on accurate identification of the flow stress model. Plastometric tests were performed for each material from the database and the

flow stress models were identified using the inverse analysis (Szeliga et al. 2006). The relation between the flow stress (σ_p) and the average pressure (p_{av}) in rolling has to account for so called friction hill and this relation was described by the surface response method. The metamodel follows the idea of Sims (Sims 1954), who introduced a coefficient Q representing the average pressure-to-flow stress ratio ($Q = p_{av}/(a\sigma_p)$, where $a = 2/\sqrt{3}$). Large number of calculations was performed using FE program (Pietrzyk 2000) for different reductions (ε), roll radius (R) and friction coefficients (μ). This data were used to find polynomial relation describing the function $Q = f(\varepsilon, R, \mu)$. However, the sensitivity analysis has shown that the effect of the design variables can be combined together by introduction of one variable $\xi = \mu/\Delta$, where Δ is the shape factor defined as h_{av}/l_d , $h_{av} = (h_1 + h_2)/2$ is an average thickness, $l_d = \sqrt{Rh_1\varepsilon}$ is the length of the arc of contact. Several FE simulations were performed for various process parameters and it was found in (Szeliga et al. 2011) that for a wide range of strip thicknesses and reductions the relationship between Q and ξ is linear. In consequence, the following equation was obtained by approximation of results of the FE simulations:

$$F = \sigma_p l_d w \left(1 + 0.572 \frac{\mu}{\Delta} \right) \quad (1)$$

where: w - width of the strip, F - rolling force.

Several flow stress models are implemented in the system, described in (Pietrzyk et al. 2015) and not reported here. The Hensel-Spittel model (Hensel and Spittel 1979) was used in the case study presented in subsequent section.

$$\sigma_p = A\varepsilon^B e^{-C\varepsilon} \dot{\varepsilon}^D e^{-ET} \quad (2)$$

where ε - strain, $\dot{\varepsilon}$ - strain rate, T - temperature in $^{\circ}\text{C}$, A , B , C , D , E - coefficients. Similar metamodel was developed to calculate strain distribution through the thickness of the strip. Numerous FE simulations were performed for various parameter and it was found that following function describes strain distribution with good accuracy:

$$\varepsilon(y) = \frac{2}{\sqrt{3}} \ln \left(\frac{h}{h_1} \right) \left[1 + 3 \left(\frac{0.387y\Delta}{y_{max}} \right)^2 \right] \quad (3)$$

where: h_1 , h - entry and current thickness of the strip, respectively, y - coordinate through the thickness ($y = 0$ in the center and $y = y_{max} = h/2$ at the surface). Equation (3) allows to calculate strain at each location along the roll gap and through the thickness of the strip. Equation (1) allows further to calculate rolling torque (M_r) as well as electric current (I) and power (P) of the motor:

$$M_r = \psi F l_d \quad (4)$$

$$M_b = 2\mu_b F R_b \quad (5)$$

$$M_M = \frac{M_w + M_b + M_{bj}}{\eta_T \eta_M i} \quad (6)$$

$$P = M_s \omega \quad (7)$$

where: ψ - lever arm of the torque according to (Roberts 1983), i - transmission ratio, M_b - friction torque in bearings, M_N - nominal torque of the motor, M_{bj} - idle torque assumed as equal to $0.05M_N$, R_b - radius of the bearing, μ_b - friction

coefficient in the bearing, η_T , η_M - current and nominal angular velocity of the motor.

The equations describing components of the total torque were taken from [8]. The electric current of the motor was calculated from the electric power, depending on the type of the motor. For alternative current motors it was:

$$I = \frac{P}{U \cos(\varphi)} \quad (8)$$

where: U - voltage, φ - phase angle between voltage and current.

The equations presented in this section are used in the system to calculate all mechanical parameters as well as power and electric current required to run the rolling mill. These equations are coupled with thermal and microstructural models.

Thermal and Microstructural Models.

One dimensional solution of the Fourier equation was used to calculate temperature distribution through the thickness:

$$\frac{\partial}{\partial x} \lambda \frac{\partial T}{\partial x} + Q = \rho c_p \frac{\partial T}{\partial t} \quad (9)$$

where: T - temperature, λ - conductivity, x - coordinate along the thickness, ρ - density, c_p - specific heat, t - time.

Equation 9 has to satisfy the boundary condition on the top and bottom surface:

$$\lambda \frac{\partial T}{\partial x} = \alpha(T - T_a) \quad (10)$$

where: T_a - ambient temperature, α - heat transfer coefficient selected according to the current location of the strip.

Microstructure evolution model was based on works of Sellars (Sellars 1979). This model includes equations describing recrystallization and grain growth. The model is executed depending on strain and recrystallized material fraction according to the diagram presented in Figure 3. The diagram contains general approach for the austenite microstructure evolution simulation during the hot rolling process.

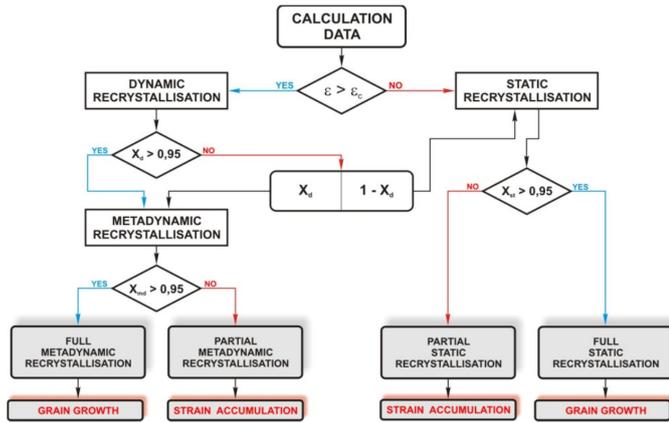


Figure 3: Flow of Calculations in Simulations of Microstructure Evolution (SRX, DRX and MTDRX Static, Dynamic and Metadynamic Recrystallization, respectively)

Prediction of the kinetics of phase transformations and volume fractions of phases after cooling was the main task of

the laminar cooling model. Modified Johnson-Mehl-Avrami-Kolmogorov (JMAK) equation was used to reach this task. The basic form of this equation is:

$$X = 1 - e^{-kt^n} \quad (11)$$

where: X - volume fraction of a new phase, k , n - coefficients, t - time. Modification of equation 11 included introduction of the coefficient k as a function of the temperature. Various functions $k = f(T)$ were used for different transformations (Pietrzyk and Kuziak 2012).

Coefficients in microstructure evolution equations for each steel in the database of the system were determined on the basis of stress relaxation tests or 2-step compression tests. Equations of the microstructural model with the description of coefficients for investigated steels are given in (Kuziak and Pietrzyk 2011). 1D thermal model allows to obtain very fast and reliable calculations, which do not require parallelization. On the other hand, microscale models are attached to selected integration points in hierarchical semi- or fully coupled way. Such implementation offers very flexible way of distribution of calculations onto HPC infrastructures, as well as facilitates collection of results. The algorithms used in microscale can be designed and implemented on the basis of different approaches, depending on the needs. This gives another possibility of distribution on the level of multicore computing devices e.g. GPGPUs or dedicated co-processors. Additionally, the computational module contains two numerical libraries: i) optimization library dedicated to determine optimal parameters for rolling mill devices; ii) sensitivity analysis library allowing investigation of particular parameters influence on final properties of the product.

EXPERIMENTAL EVALUATION

Rolling Mill Design

The created system was validated for typical configuration of the hot rolling line consisting of furnace, descaler, roughing mill, finishing mills, laminar cooling and coiler. The configuration of the most important parameters of particular devices is given in Table 1. The material model and its parameters are given in Table 2.

Table 1: Configuration Parameters of Devices

Device	Parameter	Value
Furnace	Temperature	1250 C°
	Slab width	1.5 m
	Slab length	11 m
	Slab thickness	0.22 m
Descaler	Pressure	180MPa
	Distance	0.3m
Roughing	Reversing mill with 5 passes	0.183, 0.141, 0.099, 0.058, 0.034 m
	Linear velocities	1.5, 1.8, 2.4, 3.0, 5.0 m/s
Finishing	6 two-high roll stands	0.023, 0.014, 0.0094, 0.0067, 0.0049, 0.004 m
	Linear velocities	1.17, 1.94, 2.97, 4.28, 5.88, 7.23 m/s
Laminar cooling	8 long sections divided into 2 subsections	Intensive cooling (1st subsection) = 30 m Normal cooling (1st subsection) = 30 m Intensive cooling (2nd subsection) = 40 m Normal cooling (2nd subsection) = 20 m
	Coiler	Distance from laminar cooling

Table 2: Material Models and their Parameters

Parameter	Value
Chemical composition	"Mn": 0.7, "C": 0.16, "Si": 0.03, "Nb": 0.01
Rheological model:	"A": 2481.1, "B": 0.083937, "C": 0.1, "D": 0.12065,
Hensel-Spittel	"E": 2.9803
Phase logic params	"Ca1": 0.0349, "Ca2": -0.0000403, "Cga1": 4.57,
	"Cga2": -0.005412, "Cgb1": -0.94, "Cgb2": 0.00228

For the evaluation purpose, we designed two data farming experiments. The input parameter set for both experiments contained the pressure parameter in 3 out of 8 devices performing laminar cooling. The parametrized devices were selected based on their more important role in the process. The pressure parameter is a float number with minimum value of 0 and maximum value of 100. The factorial design with 6 values for each pressure parameter was used. As the result each data farming experiment consisted of 216 distinct simulation runs.

HPC Infrastructure

We used two TOP500 clusters (Prometheus with peak performance of 2.4 PFlop/s and Zeus with peak performance of 374 TFlop/s) available at the Academic Computer Center Cyfronet AGH as our testbed. Prometheus includes Intel Haswell cores and NVidia K40XL GPU, while Zeus consists of Intel/AMD cores supported by NVidia M2050/M2090 GPU. Computing jobs simulating the metallurgical process were scheduled to both clusters through the Scalarm platform. Each scheduled job on Prometheus had been configured to use 24 cores, and each job scheduled on Zeus had been configured to use 12 cores. In both cases we scheduled 25 jobs on each cluster for the demonstration purpose.

Results

The collected results from the conducted evaluation can be split into two groups: domain-specific results and execution-related metrics.

The first group of results, domain-specific, includes information how the temperature of the rolling strips was changing in different places throughout the simulated metallurgical process. An example of such changes from a single simulation run is presented in Figure 4. Due to asymmetric nature of the results the temperature was measured at three points on the sample thickness: (1) axis, (2) middle and (3) surface.

Another domain-specific knowledge obtained from the simulations is information about volume fraction of phases with data collected on the surface part only. This information is essential to understand and predict the kinetics of phase transformations. Sample results are depicted in Figure 5.

In Table 3 the runtime metrics of simulation runs are collected from the experiments. The total experiment runtime includes runtime of all simulation runs including the queuing time. The executed numerical simulations are thread-based parallel applications. The observed runtime decrease is due to much higher cores-per-node density on Prometheus comparing to Zeus (24 to 12) and much more efficient cores themselves. However, the executed simulation is not linearly scalable between clusters. In addition, the queuing time was lower on

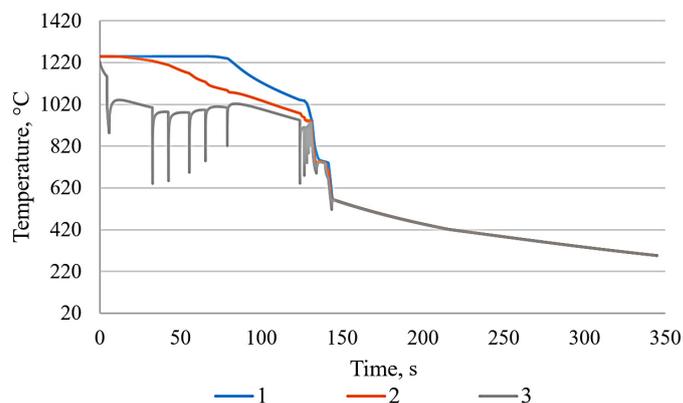


Figure 4: Temperature Change of the Simulated Steel Strips Throughout the Rolling Process (Measurement Points: 1 - Axis; 2 - Middle; 3 - Surface)

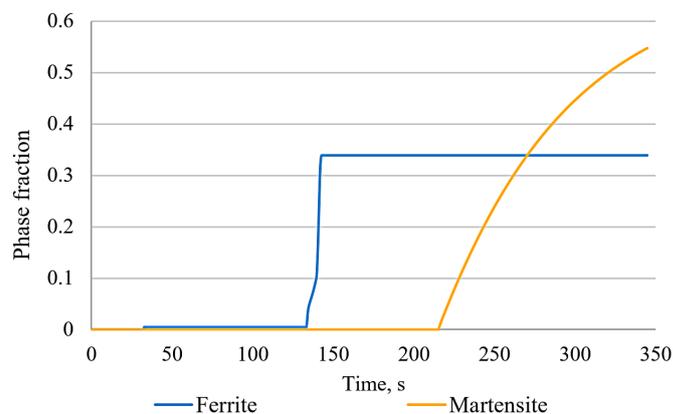


Figure 5: Simulated Volume Fractions of Phases

Prometheus, which is reflected by the total experiment runtime metric.

Table 3: Execution Time of Simulation

Cluster	Mean simulation runtime [min]	Total experiment runtime [min]
Zeus	53	606
Prometheus	38	392

CONCLUSIONS AND FUTURE WORK

In this paper we described the VirtRoll simulation system integrated with HPC clusters through the Scalarm platform for large-scale massive parallel computations. The system supports design of production technology of hot rolling strips. The main advantages of the VirtRoll system in comparison to other existing metallurgical systems is capability to study different numerical models, describing material behavior under loading conditions or during heat treatment. This allows to investigate the whole production process regarding aspects of data uncertainty and influence of different parameters of the models on final results of numerical simulations.

Detailed analysis of the results will be conducted in our future work. We plan to utilize data exploration techniques including sensitivity analysis and optimization methods, which can be

applied as the last step of the described experimental workflow. The main objective of this step is to develop better understanding of the hot rolling process and to discover pros and cons of the designed virtual rolling mill in terms of thermo-mechanical properties of the final products. In this context, data farming experiments will be used to provide large enough volume of data to perform a meaningful sensitivity analysis in order to reveal hidden relationships between simulation input and output.

ACKNOWLEDGEMENTS

Financial support of the Research Fund for Coal and Steel (RFCs), project no. RFSR-CT- 2013-00007, is acknowledged.

REFERENCES

- Funika, W.; P. Żak; and G. Łaganowski. 2015. "IEMAS-Aided Exploration of Sensitivity Analysis Methods Implemented in MATLAB and R". In *Proceedings of CGW15*, M. Bubak; M. Turala; and K. Wiatr (Eds.). ACC-Cyfronet AGH, 71–72.
- Hensel, A. and T. Spittel. 1979. *Kraft- und Arbeitsbedarf Bildsamer Formgebungs-verfahren*. VEB Deutscher Verlag für Grundstoffindustrie, Leipzig.
- Horne, G. and S. Seichter. 2014. "MSG-088 Data Farming in Support of NATO". Technical report. <http://ftp.rta.nato.int/public/PubFullText/RTO/MP/STO-MP-MSG-111/MP-MSG-111-14.pdf>.
- Król, D. and J. Kitowski. 2016. "Self-scalable Services in Service Oriented Software for Cost-effective Data Farming". *Future Generation Computer Systems* 54, 1–15.
- Król, D.; M. Orzechowski; J. Kitowski; C. Niethammer; A. Sulisto; and A. Wafai. 2014. "A Cloud-Based Data Farming Platform for Molecular Dynamics Simulations". In *Proc. of 7th IEEE/ACM Int. Conf. UCC*, IEEE, London UK, 579–584.
- Król, D.; R. Słota; J. Kitowski; Ł. Dutka; and J. Liput. 2014. "Data Farming on Heterogeneous Clouds". In *Proc. of IEEE 7th Int. Conf. CLOUD*, 873–880.
- Kuziak, R. and M. Pietrzyk. 2011. "Physical and Numerical Simulation of the Manufacturing Chain for the DP Steel Strips". In *Steel Research International, Int. Conf. ICTP*, Aachen, 756–761.
- Kvassay, M.; L. Hluchy; S. Dlugolinsky; M. Laclavk; B. Schneider; H. Bracker; A. Tavcar; M. Gams; D. Król; M. Wrzeszcz; and J. Kitowski. 2012. "An Integrated Approach to Mission Analysis and Mission Rehearsal". In *Proc. of the Winter Simulation Conference, WSC 12*, 362:1–362:2.
- Pietrzyk, M. 2000. "Finite element simulation of large plastic deformation". *Journal of Materials Processing Technology* 106, 223–229.
- Pietrzyk, M. and R. Kuziak. 2012. "Modelling Phase Transformations in Steel". In *Microstructure Evolution in Metal Forming Processes*, J. Lin; D. Balint; and M. Pietrzyk (Eds.). Woodhead Publishing, Oxford, 145–179.
- Pietrzyk, M.; L. Madej; L. Rauch; and D. Szeliga. 2015. *Computational Materials Engineering: Achieving High Accuracy and Efficiency in Metals Processing Simulations*. Elsevier, Amsterdam.
- Rauch, L. 2012. "Hybrid Computer System for the Design of Flat Rolling Technology Case Study for Multiphase Steel". *Computer Methods in Materials Science* 12, No.3, 218–224.
- Rauch, L.; R. Gotab; R. Kuziak; V. Pidvysotsky; and M. Pietrzyk. 2012. "Application of the Expert Computer System to Design the Hot Strip Rolling Technology in the LPS Line". *Transactions of the Instytut Metalurgii Żelaza* 64, No.1, 104–109.
- Rauch, L.; D. Szeliga; D. Bachniak; K. Bzowski; R. Słota; M. Pietrzyk; and J. Kitowski. 2015. "Identification of Multi-Inclusion Statistically Similar Representative Volume Element for Advanced High Strength Steels by Using Data Farming Approach". *Procedia Computer Science* 51, 924–933.
- Roberts, W. 1983. *Hot Rolling of Steel*. Marcel Dekker, Inc., New York.
- Sellars, C. 1979. "Physical Metallurgy of Hot Working". In *Hot Working and Forming Processes*, C. Sellars and G. Davies (Eds.). The Metals Soc., London, 3–15.
- Sims, R. 1954. "The calculation of Roll Force and Torque in Hot Rolling Mills". *Proceedings of the Institution of Mechanical Engineers* 168, 191–200.
- Szeliga, D.; J. Gawad; and M. Pietrzyk. 2006. "Inverse Analysis for Identification of Rheological and Friction Models in Metal Forming". *Computer Methods in Applied Mechanics and Engineering* 195, 6778–6798.
- Szeliga, D.; L. Sztangret; J. Kusiak; and M. Pietrzyk. 2011. "Two Approaches to Identification of the Flow Stress Model Application of the Metamodel". In *Proc. XXX Verformungskundliches Kolloquium*, Planneralm, 109–114.

AUTHOR BIOGRAPHIES

Dariusz Król obtained M.Sc and Ph.D. in computer science. His fields of interest include distributed, grid and cloud computing, software engineering, storage systems, data farming, scalability. E-mail: dkrol@agh.edu.pl; Web-page: <http://www.ki.agh.edu.pl/en/staff/krol-dariusz>.

Renata Słota obtained M.Sc. and Ph.D. in computer science. Her fields of interest include storage systems, parallel and distributed computing, development of grid and cloud systems. Autor of over 150 publications. Participant in many international and national projects. E-mail: rena@agh.edu.pl; Web-page: <http://www.ki.agh.edu.pl/en/staff/slota-renata>.

Jacek Kitowski full professor of computer science, Head of Computer Systems Group at Department of Computer Science AGH and Head of International Affairs at ACK Cyfronet AGH. His interests include high performance computing, grid and cloud environments and knowledge engineering. Director of PLGrid consortium for development of national computing infrastructure for scientific research in Poland. Autor of over 250 publications. Evaluator of grants and papers. E-mail: kito@agh.edu.pl; Web-page: <http://www.icsr.agh.edu.pl/~kito>.

Łukasz Rauch obtained M.Sc. and Ph.D. in computer science. His fields of interest include heterogeneous computing, multiscale modelling, pattern recognition. Involved in many international and national grants and collaborations. E-mail: lrauch@agh.edu.pl; Web-page: <http://home.agh.edu.pl/~lrauch>.

Krzysztof Bzowski works at AGH University (Faculty of Metals Engineering and Industrial Computer Science). Interests include statistical representations of materials microstructure and numerical simulations. E-mail: kbzowski@agh.edu.pl; Web-page: <http://home.agh.edu.pl/~kbzowski>.

Maciej Pietrzyk full professor at Faculty of Metallurgy and Material Sciences AGH, Head of the Department of Applied Computational Science and Modelling since 1997. Scientific interest focuses on application of numerical methods in metallurgy and materials science. Author of over 500 publications, co-author of two books published by Springer and Elsevier. Evaluator of grants and papers. E-mail: Maciej.Pietrzyk@agh.edu.pl; Web-page: <http://home.agh.edu.pl/~mpietrz>.

FUTURE DEMAND UNCERTAINTY IN PERSONNEL SCHEDULING: INVESTIGATING DETERMINISTIC LOOKAHEAD POLICIES USING OPTIMIZATION AND SIMULATION

Michael Römer

Institute of Information Systems and Operations Research
Martin Luther University Halle-Wittenberg, Germany
Email: michael.roemer@wiwi.uni-halle.de

Taïeb Mellouli

Institute of Information Systems and Operations Research
Martin Luther University Halle-Wittenberg, Germany
Email: mellouli@wiwi.uni-halle.de

KEYWORDS

Personnel Scheduling; Uncertainty, Optimization, Simulation; Computational Stochastic Optimization; Lookahead Policies

ABSTRACT

One of the main characteristics of personnel scheduling problems is the multitude of rules governing schedule feasibility and quality. This paper deals with an issue in personnel scheduling which is both relevant in practice and often neglected in academic research: When evaluating a schedule for a given planning period, the scheduling history preceding this period has to be taken into account. On the one hand, the history restricts the space of possible schedules, in particular at the beginning of the planning period and with respect to rules a scope transcending the planning period. On the other hand, the schedule for the planning period under consideration affects the solution space of future planning periods. In particular if the demand in future planning periods is subject to uncertainty, an interesting question is how to account for these effects when optimizing the schedule for a given planning period. The resulting planning problem can be considered as a multistage stochastic optimization problem which can be tackled by different modeling and solution approaches. In this paper, we compare different deterministic lookahead policies in which a one-week scheduling period is extended by an artificial lookahead period. In particular, we vary both the length and the way of creating demand forecasts for this lookahead period. The evaluation is carried out using a stochastic simulation in which weekly demands are sampled and the scheduling problems are solved exactly using mixed integer linear programming techniques. Our computational experiments based on data sets from the Second International Nurse Rostering Competition show that the length of the lookahead period is crucial to find good-quality solutions in the considered setting.

INTRODUCTION

Personnel scheduling problems have been widely studied in the Operations Research literature. The interest in this class of problems stems from the fact that, compared to scheduling problems dealing with more “simple” resources, personnel-related problems are considerably more complex: In personnel scheduling problems, a multitude of rules and objectives need to be considered. An important source of these rules is the human need for daily, weekly and annual recreation. Furthermore, quality-of-life aspects often play an

important role in the objective function of personnel scheduling problems: People value having a full weekend off; furthermore, they have individual preferences for days and shifts off and other schedule attributes. On the one hand, these individual preferences lead to the fact that there are often different contracts such as full-time and part-time contracts and sometimes even different payment schemes. On the other hand, personalized preferences expressed e.g. via requests for days off lead to the fact that personnel scheduling problems often cannot simply be regarded on a resource-aggregated level but every staff member needs to be considered individually.

While the challenging characteristics of personnel scheduling problems sketched so far form a good source of challenges for the scientific community in the field of Operations Research – for a recent overview of the research dealing with personnel, see e.g. (Van den Bergh et al., 2013) – many of those studies deal with simplified and stylized problem instances. This paper deals with two practically relevant issues to be considered in real-world personnel scheduling often neglected in academic papers: The multistage nature of personnel scheduling problems and the uncertainty of the demand to be covered in future planning periods. In practice, when evaluating a schedule for a given planning period, the scheduling history preceding this period has to be taken into account. On the one hand, this restricts the space of possible schedules, in particular at the beginning of the planning period and with respect to rules a scope transcending the planning period. On the other hand, the schedule for the planning period under consideration affects the solution space of future planning periods. In particular if the demand in future planning periods is subject to uncertainty, an interesting question is how to account for these effects when optimizing the schedule for a given planning period.

In this paper, we show that the planning problem resulting from considering the described multistage characteristic along with uncertain future demands can be considered as a multistage stochastic optimization problem. Furthermore, we propose deterministic lookahead policies for this problem in which at each stage, a mixed-integer linear programming problem is solved for the planning period under consideration augmented with a lookahead period. Finally, we evaluate these policies using publicly available problem instances from the Second International Nurse Rostering Competition (INRC-II).

The paper is structured as follows: In the next section, we describe the problem setting of the INRC-II forming the source of the data sets used in the computational experiments. Then, we provide a short review of related work followed

by a characterization of the personnel scheduling problem considered in this paper as a multistage stochastic optimization problem. Next, we present the deterministic lookahead policies to be investigated in this paper followed by the description of the experimental design and the computational results.

PROBLEM SETTING

The problem setting along with the data sets considered in this paper stems from the Second International Nurse Rostering Competition (INRC-II). For a detailed description of this competition and its problem setting, see (Ceschia et al., 2015). In this section, we will briefly sketch its main characteristics relevant to our investigation.

The INRC-II problem consists in finding a cost-minimal schedule (a sequence of shift assignments and days off) for a given set of nurses covering the given shift-wise demand and respecting all hard roster legality rules for a given planning horizon. The cost function consists of a linear combination of penalties for violating soft rules of the problem. Furthermore, each nurse has a given set of skills and certain contract (e.g. full time, half time and part time) governing certain rule-related parameters such as the maximum number of working days in the planning horizon. Using the numbering and notation from (Ceschia et al., 2015) in which **H** stands for a hard and **S** for a soft constraint, the constraints used for evaluating a solution are as follows:

H1 A nurse can be assigned at most one shift per day.

H2 For each (day/shift/skill) combination, the assigned number of nurses must cover the minimum requirement.

H3 Two consecutive shifts of one nurse must form a legal shift type succession (e.g., early must not follow night shift)

S1 Respect the the optimal requirement for each (day/shift/skill) combination

S2 Respect the minimum and maximum number of consecutive work days (in general and for each shift type)

S3 Respect the minimum and maximum number of consecutive days off

S4 Respect the shift off requests for each nurse

S5 At a weekend, a nurse should either work both days or no day at all

S6 Respect the minimum and maximum number of total work days in the planning horizon

S7 Respect the maximum number of total working weekends.

Since they affect the legality of blocks of days on and days off, in the following discussion, we will refer to the rules H3, S2, and S3 as “block-related” rules. Similarly, since they affect the full planning horizon, we refer to the rules S6 and S7 as “full-horizon” rules.

One of the main features of INRC-II competition is the fact that the problem to be solved is a multistage problem under uncertainty: While each instance consists of a four- or eight-week scheduling horizon, demand and request information only becomes available for a single week in each stage. Thus,

in each stage, a single-week scheduling problem needs to be solved – under consideration of the history from the previous week(s) affecting the evaluation of the full-horizon rules and of the block-related rules at the beginning of the week. Given the last statement, it becomes clear that a main challenge of the described problem lies in both finding a good schedule for the week under consideration and leaving a history allowing finding good schedules for the subsequent week(s) – for which both demand and request information is unknown. Following Powell (2014), this type of problem can be considered as a multistage stochastic resource allocation problem.

Each publicly available data set for the INRC-II consists of multiple files: A scenario file containing nurse-, contract- and rule-related data, multiple history files which can serve as a history for the first week and 10 week data files containing demand and preference information. Note that in order to reduce the computational burden for the extensive experiments performed for the present paper, for each of the considered INRC-II instances, we only consider the skill “trainee”. On the one hand, this reduces the number of nurses to be regarded per instance. On the other hand, since in all instances, the trainee nurses only have a single skill, the originally multi-skill setting of the INRC-II problem is turned into a single-skill setting.

RELATED WORK

For a recent and comprehensive survey of the Operations Research literature dealing with variants of personnel scheduling problems, see (Van den Bergh et al., 2013). This section intends to provide a short overview of work closely related to the problem sketched in the previous section and to our approaches used in the following sections.

When it comes to solving (deterministic) personnel scheduling problems, as discussed by Van den Bergh et al. (2013), one can distinguish exact and heuristic methods. Many of the most successful exact approaches are based on Mathematical Programming, making use of state-of-the art solvers complemented by problem-specific valid inequalities (see e.g. Santos et al., 2014) and/or advanced techniques such as branch-and-price (see e.g. Burke and Curtois, 2014). Note that while we use a Mixed-Integer Linear Programming approach to solve the scheduling problem in each stage, the focus of the present paper is on evaluating policies for handling the multistage stochastic nature of the problem. Consequently, besides the fact that we use an exact approach for solving these problems in order to avoid issues with regard to solution quality introduced by heuristic approaches, the choice of the modeling and solution approach is not of primary importance for the results of the present paper.

The first important aspect considered in this paper is the multistage characteristic of the problem described in the previous section. Note that the multitude and the complexity of the schedule legality rules makes this issue particularly relevant in personnel scheduling problems: On the one hand, there are full-horizon rules transcending the planning period of each stage, on the other hand, there are local or block-related rules affecting the start of the planning period. While this issue is often ignored in the personnel scheduling literature, recently, Salassa and Vanden Bergh (2012) address this issue in a deterministic setting and show that neglecting the multistage

characteristic leads to inferior results in practice. Furthermore, it can be expected that the INRC-II competition will draw a certain amount of research interest towards this issue (which actually happened to the authors of this paper).

The second important aspect of the problem sketched in the previous section is the fact that the demand and request information in future stages is subject to uncertainty. As shown in (Van den Bergh et al., 2013), there are some works dealing with uncertainty in personnel scheduling, mostly using stochastic programming and robust optimization approaches. Most papers however, deal with a two-stage setting in which the decision stages deal with different types of decisions: For example, in (Kim and Mehrotra, 2015), the first stage involves staffing decisions and the second stage involves the selection of weekly scheduling patterns under uncertain demand. Other papers such as (Campbell, 2011) deal with the problem of creating a schedule under demand uncertainty which then has to be adjusted in an operational setting forming the second-stage subproblem. This contrasts with the problem setting described in the previous section in which the type of decisions taken at each stage have the same type: Constructing a schedule for a full week for which demand is known while demand for the subsequent weeks is subject to uncertainty. Following Powell (2014), this type of problem can be characterized as a sequential or multistage stochastic optimization problem; it consists in finding an optimal policy, that is a function mapping states to decisions. While in general, this optimization problem is intractable, approaches from the field of *approximate dynamic programming* have been successfully applied to multistage stochastic resource allocation problems, see various case studies e.g. on fleet and driver management in the less-than-truckload industry described in the monograph (Powell, 2011).

To the best of our knowledge, however, there is no study applying similar techniques to personnel scheduling problems as described in the last section. In the following sections, we demonstrate that the problem can in fact be interpreted and modeled as a multistage stochastic optimization problem. Furthermore, we propose variants of deterministic lookahead policies and evaluate these policies using stochastic simulation.

MODELING AS A MULTISTAGE STOCHASTIC OPTIMIZATION PROBLEM

The problem addressed in this paper can be considered as a (stochastic) dynamic resource allocation problem. While there are different modeling and solution frameworks for this class of problems, among which is (multistage) stochastic programming, in this paper, we will model the problem using the more general framework for multistage stochastic optimization problems discussed in (Powell, 2014). In the personnel scheduling problem under consideration, each stage t consists of a weekly scheduling problem for which the demand is assumed to be known. The future demand from stage $t + 1$ to the final stage T (in the instances considered in this paper, T is 4 or 8), however, is uncertain. Note that an interesting feature of the personnel planning problem under consideration is that each stage itself consists of a dynamic resource allocation problem (which in addition has to account for the impact of the resulting schedule on future planning weeks with uncertain demand).

According to Powell (2014), a multistage stochastic optimization can be modeled using a framework encompassing five elements, each of which we will shortly explain and apply to the personnel scheduling problem addressed in this paper.

The first element is the characterization of the **state** S_t of the system at the time t before a decision is made. S_t is the so-called state variable, which, following Powell (2014), can be defined as “the minimally dimensioned function of history that is necessary and sufficient to compute the decision function, the transition function and the contribution function”. In the case of the personnel planning problem considered here, the elements of the state variable encompass the resource state at the beginning of the planning week and the demand and request information (in the INRC-II problem, this information is supplied in form of the history and week data files). Note that in the case study under consideration, the probability distribution of demand and request data does not depend on the week index t . The resource state represents all rule-relevant information such as number of days worked so far, number consecutive work days up to the border of the planning period etc.

The next element are the **decisions** x_t to be determined by the chosen policy in stage t . In the personnel scheduling problem under consideration, the vector x_t encompasses all assignment decisions, that is, there is a binary decision variable for each combination of nurse, day d in the week t under consideration and shift type. Note that the ensuring the feasibility of the decision vector x_t is one of the things to consider when designing a policy π .

The third element of the framework is the vector of **exogenous information** W_t becoming available in period t : In case of the personnel scheduling problem, the exogenous information involves the demand for each combination of day and shift as well as the shift off requests.

The next element is the **transition function** $S^M(S_t, x_t, W_{t+1})$ describing the transition from state S_t to S_{t+1} given the decisions x_t taken in t and the exogenous information becoming available in $t + 1$. In the personnel scheduling problem considered in this paper, the transition to the resource state R_{t+1} is performed by computing the new schedule history information for each nurse based on the information R_t and on the assignment information contained in the vector x_t . Since the demand and request information is not history-dependent, the transition with regard to this part of the state variable is performed by replacing the information from period t with the newly arrived exogenous information contained in W_{t+1} .

The last element of the modeling framework is the **objective function** stating the overall objective of minimizing the expected costs over the planning horizon. Note that in the case of the planning problem under consideration, the planning horizon is finite and consists of T periods. The objective function (which is linear and does not involve a discount factor) can be formulated as follows:

$$\min_{\pi \in \Pi} \mathbb{E}^{\pi} \sum_{t=1}^T C(S_t, X_t^{\pi}(S_t)) \quad (1)$$

Note that the optimization problem (1) consists in finding the cost-minimal policy π from the set Π of all policies. Since π is a function, the problem forms a search in a function space.

DETERMINISTIC LOOKAHEAD POLICIES

At the time of this writing, no computationally tractable method for solving problem (1) exactly is known. Nonetheless, there exist methods for approximately tackling this type of problem. According to Powell (2014), there are four fundamental classes of policies (and hybrids of these classes) typically employed to address multistage stochastic optimization problems:

Policy function approximations (PFA) represent an analytic function mapping a state to an action or a set of decisions. For example, a PFA may be a simple lookup table, a decision rule or a linear or polynomial function. Note that applying a PFA does not involve solving an optimization problem.

(Myopic) Cost function approximations (CFA) are formed by modifying the objective function and/or the constraints of the decision problem to be solved in stage t in a way that the resulting policy does not only involve the single-stage cost function but also accounts for the impact of the decisions in stage t on the future stages.

Value function approximations (VFA) involve constructing and calibrating a function approximating the future value of the state resulting from taking a decision. For example, in a resource allocation problem, value function approximation can be designed around the post-decision state, that is, the state after having taken a decision, of the resources to be allocated.

Lookahead policies involve solving a multistage decision problem at each stage by explicitly considering the future decision stages for a lookahead horizon to be specified. Lookahead policies come in two main flavors: In a deterministic lookahead policy, the uncertain parameters in the exogenous information process are replaced by point estimates resulting in a deterministic multistage optimization problem to be solved in each stage. In a stochastic lookahead policy, the uncertain parameters are modeled in a stochastic scenario tree; the resulting problem to be solved then forms a (multistage) stochastic programming problem (which may be approximated by a two-stage stochastic programming problem in order to reduce the computational complexity of the problem).

For the personnel scheduling problem considered in this paper, we first tried to design myopic cost function approximations and simple value function approximations – however, we were not satisfied with the results: In fact, it was difficult to even get an intuitive understanding of what makes a “good” post-decision state (that is, the state after having taken the scheduling decisions for the week under consideration) and thus, at least to our impression, approximating the future value of a given state is difficult for this type of problem.

As a result, we started experimenting with deterministic lookahead policies which yielded much better results. This was in line with our intuition that a lookahead model would be a good approach for dealing with the impact of the block-related rules at the end of the scheduling week: Instead of explicitly stating what makes a good state at the end of a week, adding a lookahead period H may implicitly yield good end-of-the

week states since the impact of the schedule for the planning week under consideration on the subsequent week is accounted for by considering the lookahead period.

The resulting deterministic lookahead policy π , parameterized by the forecasting strategy fs and the length H of the lookahead horizon can be written as follows:

$$X_t^{\pi(fs,H)}(S_t) = \arg \min_{x_t \in \mathcal{X}_t} \left(c_{tt}x_{tt} + \sum_{t'=t+1}^{t+H} c_{tt'}x_{tt'} \right) \quad (2)$$

Using this policy, in each stage t , the optimization problem does not involve the decisions variables (and possibly needed supplementary variables) affecting t (represented by the vector x_{tt}), but also the decisions $x_{tt'}$ affecting stages contained in the lookahead period ranging from $t+1$ to $t+H$. Note that the full vector of decision variables from all these stages is denoted with x_t ; the set \mathcal{X}_t depends on the state S_t and denotes the set of all feasible vectors x_t .

Clearly, (2) forms a very high-level statement of a deterministic lookahead policy. For our experiments, we formulated (2) as a mixed-integer linear program – that is, \mathcal{X}_t is formulated as a set of linear inequalities and integer constraints – we solve using a standard solver. Note that instead of using a compact formulation as discussed in Santos et al. (2014) or an extreme-point formulation as proposed by Burke and Curtois (2014), we use a multi-commodity flow formulation in the spirit of (Mellouli, 2001) and (Römer and Mellouli, 2011) dealing with airline crew scheduling.

In order to capture the soft full-horizon constraints (in our case, concerning the limitations regarding number of days of work and of working weekends) in the model, the upper bounds belonging to the full planning horizon were adjusted according to the relative amount of time passed after the end of the lookahead horizon $t+H$.

When designing a deterministic lookahead model it is necessary to make point forecasts for the exogenous information $W_{t'}$ for the lookahead period from $t+1$ to $t+H$ for which demand is uncertain at stage t . In this paper, we consider two strategies fs for obtaining these forecasts. The first approach, referred to by $fs := R$, uses a simple resampling strategy: For each of the periods t' , one of the ten week data files in the INRC-II dataset under consideration is randomly drawn and the demand information from this week data file is used as a forecast.

The second approach is based on (conditional) averaging and referred to by $fs := A$ in the experimental results section. Based on an analysis of the weekly demand data of all INRC-II data sets, we figured out that in most cases, the demand of weekdays is significantly different from the demand on weekends. As a result, for each data set, we compute the average demand for weekdays and weekend days across all week data files and use these average values as point forecasts for the respective types of days (the weekday average is thus use from Monday to Friday and the weekend average for Saturday and Sunday).

In addition to varying the forecasting strategy, we experiment with the length of the lookahead horizon. Note that we

do not only use full week increments, but also consider a lookahead horizon of three days. In this case, the H takes the fractional value $3/7$ and the lookahead model is constructed accordingly.

EXPERIMENTAL DESIGN AND RESULTS

The problem instances for evaluating the policies described above are publicly available and stem from the Second International Nurse Rostering Competition (INRC-II) described in (Ceschia et al., 2015). As discussed in the problem description above, in order to manage the computational burden of the experiments, we artificially reduced these instances by only considering trainee nurses, resulting in a reduction with respect to the number of nurses to approximately a quarter of the original number (the tables below display the number of trainee nurses per instance).

For each of the resulting instances, we evaluate the performance of the policies described in the previous section by a simulation-based approximation of the objective function (1): We randomly sample N paths for the full scheduling horizon T (that is, we randomly select a history file for the first week and then select a random sequence of week data files for each of the data sets). The estimated performance $\hat{F}^{\pi(f_s, H)}$ of a policy $\pi(f_s, H)$ is then estimated by applying the policy to each sample path and averaging the resulting costs:

$$\hat{F}^{\pi(f_s, H)} = \frac{1}{N} \sum_{n=1}^N \left(\sum_{t=1}^T C(S_t, X_t^{\pi(f_s, H)}(S_t)) \right) \quad (3)$$

As discussed in the previous section, in this paper, we consider deterministic lookahead policies varying with respect to two dimensions: The forecasting strategy f_s and the length H of the lookahead horizon. Concerning the forecasting strategy, we use an averaging strategy (signified by A) and a resampling strategy (signified by R). When it comes to the lookahead horizon H , we consider multiple values: First of all, we consider a myopic policy with $H = 0$ in which there is no lookahead at all in order to figure out whether a lookahead is useful at all. Then, we use a lookahead period of three days ($H = 3/7$) in order to consider a short lookahead horizon – the main motivation is to evaluate whether such a short lookahead with a comparably small computational burden is able to address the border effects resulting from block-related constraints such as shift sequence rules. Furthermore, for both the four- and the eight-week instances, lookahead horizons of 1, 2 and 3 weeks are considered (note that for a four-week instance, $H = 3$ means that in the first week, the full planning horizon of $T = 4$ is considered in the lookahead model). Moreover, for the eight-week instances, $H = 5$ and $H = 7$ are considered.

All experiments were conducted on a personal computer with 8 GB Ram, with an Intel Core i7 4-core CPU with 3.4 GHz. The simulation and the model generation are implemented in C++. For solving the mixed integer linear programs (MILP) within the deterministic lookahead approach, we used IBM ILOG CPLEX 12.6. For each instance and for each policy, we carried out $N = 1000$ replications in order to obtain a good estimate of the true value of each policy. Note that the

number of (non-trivial) MILP to solve within one replication corresponds to the number of weeks in the instance. As a consequence, since in total we had to solve 912 000 MILP, we set the mipgap parameter to 0.5 % and imposed a time limit of 4 minutes per MILP. Note, however, that this time limit was rarely ever hit.

TABLE I. AVERAGE PERFORMANCE FOR 4-WEEK INSTANCES ($N = 1000$)

# nurses	f_s	lookahead horizon H (weeks)				
		0	3/7	1	2	3
5	A	575.3	492.8	444.5	441.1	438.6
	R	575.3	494.4	447.7	443.2	439.9
8	A	614.1	613.0	559.5	544.8	532.7
	R	614.1	625.6	551.4	539.3	539.1
10	A	637.2	564.0	441.8	424.0	422.6
	R	637.2	578.8	445.6	432.8	428.9
20	A	920.3	797.3	566.5	509.9	508.6
	R	920.3	829.3	561.8	519.1	512.6
25	A	829.2	692.0	509.7	441.4	433.0
	R	829.2	680.1	514.6	449.1	443.3
30	A	1694.5	999.3	688.0	619.4	590.2
	R	1694.5	974.6	683.3	610.7	585.4

The results for the four-week instances are presented in Table I. The most evident observation concerns the effect of the lookahead horizon: For all instances (and for both forecasting strategies), increasing the lookahead horizon leads to a significant cost reduction. For all but one instance, even a three-day lookahead yields a major improvement (more than 10 % on average) compared to a myopic policy without lookahead. Then, for every additional week, the solution quality is further improved.

A very interesting result is that the positive effect of a long lookahead period grows with the size of the instances in terms of the number of nurses: The relative impact of using a long lookahead horizon is much bigger for bigger instances (e.g., the best solution is 25 % better than the myopic solution in the 5-nurse instance while it is 75 % better for the 30-nurse instance). Furthermore, the marginal positive impact of additional weeks to the horizon is higher for big instances.

Regarding the choice of the lookahead strategy, the result is not as clear: In some cases, the resampling-based strategy (R) works better, in other cases the average-based strategy (A) is more effective. Even for a single instance, one strategy does not necessarily dominate the other when considering all tested lookahead horizons. On average, however, the strategy A turns out to perform slightly better. In general, given the described results and given the fact that both strategies are fairly simple it may be beneficial to develop more sophisticated forecasting strategies in order to achieve better overall results.

Most of the statements made regarding the results for the four-week instances also hold for the results for the eight-week instances displayed in Table II. In general, a longer forecasting horizon leads to better results and in most (but not all) cases the averaging-based forecasting strategy performs better than the resampling strategy. For some instances, interestingly, even if there is some benefit of using a very long lookahead horizon, the largest part of the improvement compared to a myopic policy can be obtained with a 1- or 2-week lookahead period.

TABLE II. AVERAGE PERFORMANCE FOR 8-WEEK INSTANCES
($N = 1000$)

# n.	f_s	lookahead horizon H (weeks)						
		0	3/7	1	2	3	5	7
5	A	807.9	722.8	652.7	637.4	630.2	633.8	633.6
	R	807.9	746.8	650.5	633.7	620.6	612.8	613.2
8	A	1318.3	1572.1	1134.1	1112.6	1085.2	1065.5	1069.9
	R	1318.3	1552.2	1134.1	1106.2	1086.5	1073.0	1070.1
10	A	1374.0	1316.5	1001.8	973.2	942.6	910.3	904.2
	R	1374.0	1351.8	1012.5	973.9	944.0	913.3	909.8
20	A	1645.7	1826.7	1289.0	1194.6	1167.8	1154.3	1154.0
	R	1645.7	1832.9	1278.1	1201.5	1166.4	1153.1	1147.6
25	A	1821.6	1697.2	1204.0	1149.6	1123.8	1057.0	1040.3
	R	1821.6	1700.5	1200.2	1154.0	1120.1	1065.3	1050.7
30	A	3416.7	2082.0	1047.4	834.4	761.7	671.4	657.2
	R	3416.7	2006.0	1017.0	825.1	767.5	682.4	670.2

Again, the relative effects of the lookahead policy grow with the instance size.

Finally, it should be noted that an issue which may be considered in more detail are the characteristics of the problem instances under consideration and their effect on the choice of the policy. While the tables only show the number of nurses for describing a problem instance, the instances may vary with respect to several other factors such as the relative frequency of nurses with part-time contracts, the values of rule-related parameters such as the maximum number of consecutive work days, the demand level in relation to available nurses and the variability of demand. It can be suspected that these characteristics have a certain impact on the effectiveness of policies and may help to explain the differences in the results between the instances which cannot simply be explained by the number of available nurses.

CONCLUSIONS AND FUTURE RESEARCH OPPORTUNITIES

In this paper, we consider a class of personnel scheduling problems under future demand uncertainty in which scheduling is carried out and fixed for a planning period (e.g. one week) and demand for the future planning periods is unknown. We argue that this type of problem forms a multistage (also called sequential) stochastic optimization problem consisting in finding an optimal policy, that is an optimal decision function for each stage. Among the classes of policies which can be used to tackle this type of problem (see Powell, 2014), in this paper, we investigate variants of deterministic lookahead policies with different forecasting strategies and lookahead horizons.

The computational results impressively show that even fairly simple deterministic lookahead policies can lead to huge improvements compared to a naive myopic approach. The improvements tend to grow with the instance size with respect to the number of nurses and to the length of the scheduling horizon T as well as with the length of the lookahead horizon chosen for the lookahead policy.

The results presented in this paper form a first step to investigate personnel scheduling problems in the framework of computational stochastic optimization as discussed in (Powell, 2014). Besides trying to improve the deterministic lookahead policies investigated in this paper, a natural next step would be

to consider a stochastic lookahead policy by formulating a two- or multistage stochastic programming model. Furthermore, after thoroughly examining the problem structure and the different solutions from these policies, it might be tried to find a good value function approximation allowing to solve larger problem instances. In addition, it appears promising to develop hybrid policies, e.g. by combining a lookahead policy with a cost function approximation and/or a value function approximation.

Finally, in order allow addressing bigger problem instances, e.g. the full INRC-II instances involving all skills and multiple overlapping skill sets with up to 120 nurses, it is important to improve the modeling and solution approaches used to solve the lookahead problems.

REFERENCES

- Burke, E. K. and Curtois, T. (2014), New approaches to nurse rostering benchmark instances, *European Journal of Operational Research* 237(1), 71–81.
- Campbell, G. M. (2011), A two-stage stochastic program for scheduling and allocating cross-trained workers, *Journal of the Operational Research Society* 62(6), 1038–1047.
- Ceschia, S., Dang, N. T. T., De Causmaecker, P., Haspeslagh, S. and Schaerf, A. (2015), Second International Nurse Rostering Competition (INRC-II) - Problem Description and Rules, Technical Report arXiv:1501.04177 [cs.AI], arXiv:1501.04177 [cs.AI].
- Kim, K. and Mehrotra, S. (2015), A Two-Stage Stochastic Integer Programming Approach to Integrated Staffing and Scheduling with Application to Nurse Management, *Operations Research* 63(6), 1431–1451.
- Mellouli, T. (2001), A Network Flow Approach to Crew Scheduling based on an Analogy to a Train/Aircraft Maintenance Routing Problem, S. Voss and J. Daduna, eds, *Computer-Aided Scheduling of Public Transport*, Vol. 505 of *LNEMS*, Springer, Berlin, pp. 91–120.
- Powell, W. B. (2011), *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, John Wiley & Sons.
- Powell, W. B. (2014), Clearing the Jungle of Stochastic Optimization, Bridging Data and Decisions, INFORMS Tutorials in Operations Research, INFORMS, pp. 109–137.
- Römer, M. and Mellouli, T. (2011), Handling Rest Requirements and Preassigned Activities in Airline Crew Pairing Optimization, J. Pahl, T. Reinert and S. Voß, eds, *Network Optimization*, Vol. 6701, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 643–656.
- Salassa, F. and Vanden Berghe, G. (2012), A stepping horizon view on nurse rostering, D. Kjenstad, E. K. Burke and B. McCollum, eds, *Proceedings of the 9th International Conference on the Practice and Theory of Automated Timetabling*, PATAT, Son, Norway, 28–31 August 2012, pp. 161–173.
- Santos, H. G., Toffolo, T. A. M., Gomes, R. A. M. and Ribas, S. (2014), Integer programming techniques for the nurse rostering problem, *Annals of Operations Research* pp. 1–27.
- Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E. and De Boeck, L. (2013), Personnel scheduling: A literature review, *European Journal of Operational Research* 226(3), 367–385.

OPTIMAL PRODUCTION VOLUME OF RUBBER GLOVES MOLD FOR RUBBER GLOVES PRODUCTION PLANNING

Tuanjai Somboonwiwat
Chorkaew Jaturanonda
Nattapong Chotpan
Department of Production Engineering
King Mongkut's University of
Technology Thonburi (KMUTT)
Thailand
E-mail: tuanjai.som@kmutt.ac.th

Kanogkan Leerojanaprapa
Statistics Department
King Mongkut's Institute of
Technology Ladkrabang (KMITL)
Thailand
E-mail: klkanogk@kmitl.ac.th

KEYWORDS

Production planning, Rubber glove, Rubber glove Mold.

ABSTRACT

The production process of rubber glove is the continuous production line which the hand-shaped ceramic molds are installed and dipped into the concentrated latex to form the rubber gloves. Since, the rubber gloves are diverse in terms of size, surfaces and types of latex, the multiple molds and latex are changed and used in a production process in order to produce rubber gloves from customer requirement. Thus, the changing of mold is extremely complex for production ratio which the rubber glove production planning will be employed to meet anticipate customer orders. This research therefore develops the mathematical model to find the optimal quantity of molds and rubber gloves production planning. Finally, the developed model is applied to an example data set to find the minimum total volume of rubber gloves in every mold for keeping the minimal volume of inventory.

INTRODUCTION

Rubber gloves industry has become increasing significant for Thai economy. At present, a large volume of natural latex is used for producing rubber gloves which extracted from rubber trees. Rubber is one of the important industrial crops for Thailand, both for the local consumption and the global market. From the export data of Thai Custom Department in 2014 it was found that rubber gloves had contributed an enormous export values to Thai economy. The uses of gloves as part of infection control within healthcare will be increased extensively in the near future, as at present people are more and more aware of their health and cleanliness. The growth of sale will also affect the manufactures to serve their customer needs.

Rubber gloves industry has a complicate and complex process in production planning due to several of type and size of products. Moreover, each type of glove uses different raw materials to produce. The process of producing rubber gloves is forming by dipping each size

hand-shaped ceramic mold into a tank full of concentrated latex with the length of time. Moreover, the machines will be run all the time. The volume of production in each batch will depend on a number of hand-shaped ceramic molds which are set up within each machine by size and type of mold. Therefore, in production planning process, it is obligatory to find a number of hand-shaped ceramic molds which are fit with orders from customers. Excess or shortage items should be minimal in each type and size of production run. If the production planning process does not match with customers' orders, it is necessary to stop the production run in order to set up the machine again by changing molds. This will result in losing time and budget in setting up the machines.

Most of the literature in the area of production management for the rubber glove industry is related to the structure of the rubber industry (Haan et al. 2003), production improvement (Jirasukprasert et al. 2012) and economics and environment (Rattanapana et al. 2012). The researches on production planning for the rubber gloves manufacturing are still limited. Klomsae et al. 2012 present an applied mathematical model for rubber latex industry decision planning, that covers purchase and storage over a multi-period timeframe, with due consideration of product aging and deterioration through each time period. However, there has not been research related to the production planning of rubber glove products involving the mold. Thus, this paper presents the production planning and scheduling for rubber gloves with mixed mold sizes and types.

Most production planning and scheduling has been done by applying mathematical models (Hsu et al. 2011, Wen-Chiung et al. 2012 and Yan et al. 2013). Birger R. et. al. 2013 present the mathematical model for multi-product multi-period aggregate production-distribution planning problem with mould sharing between plants. This paper aims to create a mathematical model for rubber glove production planning and scheduling for multi-glove products, multi-molds, multi-concentrated latex and multi-production lines for multi-periods to plan production in order to meet the customer requirement.

The paper is structured as follows: In section 2, we provide general information about type of rubber gloves and rubber gloves production machine describe the production planning process. Then, the problem statement is discussed in section 3 and the mathematical model of the problem is developed in section 4. A numerical example and its computational results are presented in Section 5. Finally, concluding remarks and further work are provided in the last section.

RUBBER GLOVES INFORMATION

The types of rubber gloves, hand-shaped ceramic molds and concentrated latex are described in this section.

Types of Rubber Gloves

Generally, here are three main types of rubber glove i.e. 1) medical glove, 2) industrial glove and 3) household glove. Details are as follows:

a) Medical Glove

Medical gloves use during medical procedures to prevent contamination between caregivers and patients. Medical glove can be divided into two major types which are surgical and examination. Surgical gloves are generally sterile and feature extra long reinforced cuffs. Generally, these gloves are thicker for use in higher-risk clinical applications for extra protection. So, high technology process in production and sensitive procedures are necessary to make sure that manufacturers of these devices have a higher standard. Examination gloves are available as either sterile or non-sterile which made from 100% synthetic latex, both powder-free and powdered. Medical gloves are regulated by the Food and Drug Administration (FDA) to make sure that manufacturers can meet performance criteria such as leak and tear resistance.

b) Industrial Glove

There are various types of industrial gloves which are produced to protect against a wide variety of hazards. Normally, determining which type of glove to use is depended upon the duration of the job, the type of conditions or the environment (wet or dry). These gloves are designed to ensure employee safety and sanitary conditions in the workplace and provide both strength and chemical protection. For example, heat or cut resistant gloves have different properties and compositions from chemical and oil resistant glove.

c) Household Glove

Household glove are generally used for domestic cleaning and food processing purposes. It provides ergonomic design to ensures excellent abrasion as well as cut and tear resistance. There are different designs and colour of this type of glove in order to attract housewives. The best rubber gloves should be durable, whilst allowing users to accomplish their tasks effectively.

Hand-shaped Ceramic Molds

A number of gloves to be produced depend on a number of molds which are set up in a machine. Normally, there are many kinds of rubber glove, for instance, just smooth skin or "dot" on the fingertips or plams that let users can touch device without smudging or scratching it. Moreover, there are many sizes of rubber gloves which suit for users' hand such as XS, S, M, L, XL or XXL. In each mold conveyor, it can be mixed every size and type of rubber gloves in one production line. Hence, planners should find out the exactly molds in their factories before making a production scheduling. Table 1 shows an example of molds for each item in a case study of this research.

Table 1: An Example of Molds for Each Item

Number of hand-shaped ceramic molds						
Type\size	XS	S	M	L	XL	XXL
Smooth Skin	10,000	20,000	50,000	50,000	20,000	10,000
Dots Balm	10,000	10,000	20,000	20,000	10,000	10,000
Dots Finger	10,000	10,000	10,000	10,000	10,000	10,000

Types of Concentrated Latex

Natural rubber is obtained from different species of rubber trees in form of latex which is called normal or field latex. The latex is then placed into a centrifuge to remove some of water and increase rubber content of the latex. After centrifuging, the material is known as concentrated latex, which contains about 60% rubber, are then compounded by adding other raw material ingredients to achieve the desired performance characteristics. On the other hand, synthetic latex is produced from a petrochemical which is developed as an alternative to natural latex for some people who are allergic reaction to natural latex rubber. The 4 types of latex used in this study are Natural latex, Synthetic white latex, Synthetic blue latex, and Synthetic black latex.

PROBLEM STATEMENT

The productions planning process of rubber gloves starts from the marketing department receives the orders from customers and sends them to the production planning department for scheduling. Normally, at the first step a planner will rearrange schedule according to due date and inform a plan to other departments such as logistics and transportation department. Afterwards, the planner will check on-hand inventory and calculate actual volumes that are needed to produce. It is necessary to consider capacity of each size in all production lines in real time in order to create production schedule of each machine. If the machine is not processing the same type of rubber gloves, it will be terminated to change ceramic molds; otherwise, that orders are cancelled. After the master production

schedule (MPS) is set up, it will be returned to customers to confirm delivery time.

At present, a planner can get anticipated demand in each type and size of product in advance. A number of molds for each type and size to be fixed with each machine as well as running time process are calculated to be ensured that the products will be sent to customers on time. The amount of products depends on a number of molds which are set up at each machine by taking size and type of molds into consideration. Such process is very complicated and takes a long time for the production planner in operating and sometimes this can cause a mistake. Such optimization problem can effectively be solved using mathematical modeling. Figure 1 presents the conveyors of production lines which the molds are installed.

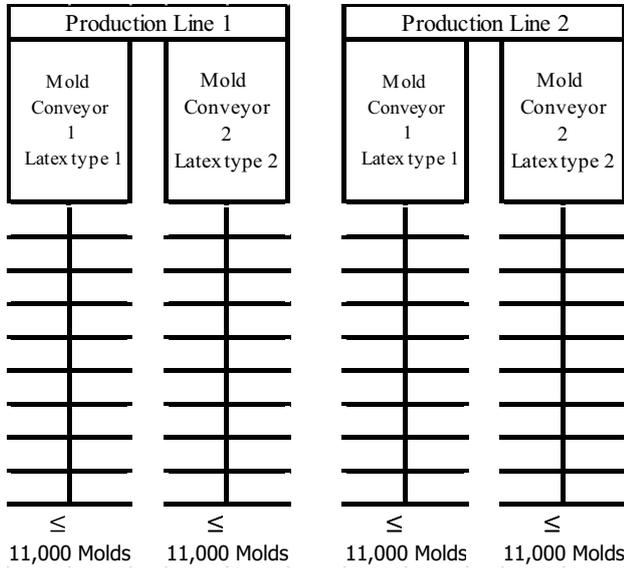


Figure 1: Mold Conveyors of Production Lines

Therefore, the production planning of rubber gloves is a process of bringing anticipated demand one month ahead to find the production volume in order to set up machines. Each machine is required to produce rubber gloves upon the customers demand. The followings are the conditions and limitation of rubber glove production machine.

- The number of mold for each machine is limited.
- Each conveyor has the same piece of mold and run simultaneously.
- Each type and size of molds is limited.
- Each machine can mix 2 types of latex. In mold conveyor 1 and mold conveyor 2 must be the same latex.

Table 2 summarizes the conditions.

Table 2: Conditions of Rubber Glove Production Machine

Production line 1	
The machine runs 24 hours per day for 30 days for each production period.	
Rubber glove mold conveyor 1	Rubber glove mold conveyor 2
<ul style="list-style-type: none"> - It can be used only one type of latex. - The maximum hand-shaped ceramic molds are less than 11,000 pieces. Therefore, the maximum capacity per day is less than $11,000 \times 26.81$ (production rate of one ceramic mold per day) or 294,910 pieces per day. - Any size of ceramic molds can be mixed, however; ceramic molds are limited. 	<ul style="list-style-type: none"> - It can be used only one type of latex. - The maximum hand-shaped ceramic molds are less than 11,000 pieces. Therefore, the maximum capacity per day is less than $11,000 \times 26.81$ (production rate of one ceramic mold per day) or 294,910 pieces per day. - Any size of ceramic molds can be mixed, however; ceramic molds are limited.

MATHEMATICAL MODEL

Details of indices, decision variables, parameters, objective functions, and constraints in production planning for rubber gloves in one month planning horizon are as follows:

Indices

- i Type and size of hand-shaped ceramic mold;
 $i = 1, 2, 3, \dots, I$
- j Product type of latex;
 $j = 1, 2, 3, \dots, J$
- k Product type mold conveyor;
 $k = 1, 2, 3, \dots, K$
- l Production line;
 $l = 1, 2, 3, \dots, L$
- t Date;
 $t = 1, 2, 3, \dots, 30$

Parameters

- d_{ijt} Demand of rubber gloves for molds ' i ' using latex ' j ' in date ' t '
- s_{ijt} Inventory of rubber gloves for molds ' i ' using latex ' j ' in date ' t '
- c_{kl} Maximum capacity per day for mold conveyor ' k ' in production line ' l '
- m_i Maximum volume of rubber gloves using molds ' i '

Decision Variables

- x_{ijklt} Quantity/Volume of rubber gloves from mold ' i ' using latex ' j ' by typing of mold conveyor ' k ' in production line ' l ' in date ' t '
- y_{jkl} Decision variable which will take on only possible states i.e. yes (1) or no (0) production by using latex ' j ', typing of mold conveyor ' k ' in production line ' l ' in date ' t '

Objective function

The objective function of the problem is to find the minimum total volume of rubber gloves in every mold 'i' and latex 'j' in 30 days, as shown in equation (1).

$$\text{Min } Z = \sum_{i=1}^I \sum_{j=1}^J s_{ij30} \quad (1)$$

where

$$s_{ijt} = x_{ijklt} \cdot y_{jkl} + s_{ijt-1} - d_{ijt} \quad ; \forall i, j, k, l, t \quad (2)$$

Constraints

1. Quantity/Volume of rubber gloves from mold 'i' using latex concentrate 'j' by typing of mold conveyor 'k' in production line 'l' in date 't' is less than or equal to maximum capacity per day for mold conveyor 'k' in production line 'l'.

$$\sum_{i=1}^I \sum_{j=1}^J x_{ijklt} \leq c_{kl} \quad ; \forall k, l, t \quad (3)$$

2. Quantity/Volume of rubber gloves from mold 'i' using latex concentrate 'j' by typing of conveyor 'k' in production line 'l' in date 't' is less than or equal to maximum volume of rubber gloves from using molds 'i'.

$$\sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L x_{ijklt} \leq m_i \quad ; \forall i, t \quad (4)$$

3. The production of rubber gloves in every mold 'i' in each conveyor belt 'k' in production line 'l' must be the same latex concentrate 'j' only.

$$\sum_{j=1}^J y_{jkl} = 1 \quad ; \forall k, l, t \quad (5)$$

4. Shortage of rubber gloves for molds 'i' using latex 'j' in date 't' is not allowed.

$$s_{ijt} \geq 0 \quad ; \forall i, j, t \quad (6)$$

5. Constraints describing the decision variables.

$$x_{ijklt} \geq 0 \quad ; \forall i, j, k, l, t \quad (7)$$

$$y_{jkl} \in \{0, 1\} \quad ; \forall j, k, l, t \quad (8)$$

The results from the production planning in each month using equation (1) to (8) will keep minimal on the volume of inventory of rubber gloves using mold 'i' in every latex 'j' in the final date of the time horizon. The schematic diagram in Table 3 shows the production quantity and demand of rubber gloves.

NUMERICAL EXAMPLE

This section demonstrates a numerical example by applying mathematical model described in section 4 using a case company in Thailand. The objective of the study is to find the number of rubber gloves volume to

be produced which keep inventory in the system at the minimal by considering the requirement of customers in each item and due date. An example of customer requirement of the rubber gloves for the smooth skin, XS size and natural latex gloves (Table 4). The first order is due on the fifteen of the month and the total requirement is 5,000,000 pieces.

Table 4: Rubber Glove Production Demand

Customer Demand				
Type	Size	Type of Concentrated Latex	Volume (piece)	Due Date
Smooth Skin	XS	natural latex	5,000,000	15
Smooth Skin	S	natural latex	5,000,000	15
Smooth Skin	M	natural latex	5,000,000	15
Smooth Skin	L	natural latex	5,000,000	15
Smooth Skin	XL	natural latex	5,000,000	15
Smooth Skin	XXL	natural latex	5,000,000	15
Smooth Skin	XS	natural latex	2,000,000	20
Dots Balm	L	natural latex	5,000,000	20
Dots Balm	XL	natural latex	5,000,000	20

The results of production planning and scheduling of rubber gloves will be displayed in terms of decision variables of the volume of gloves to be produced using mold 'i' latex 'j' for mold conveyor 'k' on production line 'l' at date 't'. An example of on hand inventory of rubber gloves is zero inventory in first period. All data will be computed under production ratio of a machine using Excel solver. The results of the volume of molds to be set up for a mold conveyor 'k' on production line 'l' at date 't' are demonstrated in Table 5. By applying these formula in a production planning procedure and adjusting mold by following computational results would result in the total of inventory volumes for every mold and latex in 30 days at the minimum.

CONCLUSION AND FURTHER STUDY

This paper, the mathematical model was formulated to minimize total volume of rubber gloves under limited resource constraints. A numerical example of case study was presented. The short-term scheduling time horizon for rubber gloves production is one month, however; this model can be extended for multi-stage decision making in the future. The proposed production planning and scheduling model can improve and organize several management decisions for producing, carrying inventories, and balancing utilization of machines in the short-term planning time horizon. However, for analysis process, it may need person who can understand mathematical precisely and take time to solve such problem. Consequently, developing decision support system (DSS) can assist and support decision makers in accommodating changes in the input information.

Table 3: The Production Volume Addressed in Mathematical Model for Rubber Glove Production Planning

Rubber Gloves Production Volume of Mold i , Latex j , Mold Conveyor k , and Production Line l , in Date t (piece) X_{ijklt}								Rubber Gloves Demand of Mold i , Latex j , in Date t (piece) D_{ijt}				
Mold	Latex	Mold Conveyor	Production Line	Date of Production				Date of Rubber Glove Demant				
				t				t				
i	j	k	l	1	2	...	30	1	2	...	30	
1	1	1	1	X_{11111}	X_{11112}	...	X_{111130}	D_{111}	D_{112}	...	D_{1130}	
			2	X_{11121}	X_{11122}	...	X_{111230}					
			:					
		1	X_{11111}	X_{11112}	...	X_{111130}						
		2	1	X_{11211}	X_{11212}	...	X_{112130}					
			2	X_{11221}	X_{11222}	...	X_{112230}					
	:							
	:	:	:	:	:	:	:	
	2	1	1	1	X_{12111}	X_{12112}	...	X_{121130}	D_{121}	D_{122}	...	D_{1230}
				2	X_{12121}	X_{12122}	...	X_{121230}				
				:				
		2	1	1	X_{12211}	X_{12212}	...	X_{122130}				
2				X_{12221}	X_{12222}	...	X_{122230}					
:								
:	:	:	:	:	:	:		
:	:	:	X_{12k1}	X_{12k2}	...	X_{12k30}		
2	1	1	1	X_{21111}	X_{21112}	...	X_{211130}	D_{211}	D_{212}	...	D_{2130}	
			2	X_{21121}	X_{21122}	...	X_{211230}					
			:					
		2	1	1	X_{21211}	X_{21212}	...					X_{212130}
				2	X_{21221}	X_{21222}	...					X_{212230}
				:
	:	:	:	:	:	:	:	
	2	1	1	1	X_{22111}	X_{22112}	...	X_{221130}	D_{221}	D_{222}	...	D_{2230}
				2	X_{22121}	X_{22122}	...	X_{221230}				
				:				
		2	1	1	X_{22211}	X_{22212}	...	X_{222130}				
				2	X_{22221}	X_{22222}	...	X_{222230}				
:								
:	:	:	:	:	:	:		
:	:	:	X_{22k1}	X_{22k2}	...	X_{22k30}		
:	:	:	:	
				X_{2jk1}	X_{2jk2}	...	X_{2jk30}	D_{2j1}	D_{2j2}	...	D_{2j30}	
i	j	k	l	X_{ijk1}	X_{ijk2}	...	X_{ijk30}	D_{ij1}	D_{ij2}	...	D_{ij30}	

Table 5: Mold and Type of Concentrated Latex Installation Result

Mold and Latex Install Result (Piece)																						
Rubber Glove Production Machine		Smooth Skm Mold						Dots Balm Mold						Dots Finger Mold								
Production Line	Mold Conyenyer	Type of Concentrated Latex	XS	S	M	L	XL	XXL	XS	S	M	L	XL	XXL	XS	S	M	L	XL	XXL		
1	1	Natural latex	1,467	2,933	733		1,467	6,600						1,467								
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				
1	2	Natural latex	1,630	1,731	2,078	1,100	1,467	4,094				1,100	1,467									
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				
2	1	Natural latex	2,933	1,467	2,933	733	1,945	1,976				733	1,945									
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				
2	2	Natural latex	3,667	2,200	4,787		1,100	1,813						1,100								
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				
3	1	Natural latex	4,828	1,467	2,200	713	2,200	1,467				713	2,200									
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				
3	2	Natural latex	3,300	2,933	5,093	0	1,467	1,874				0	1,467									
		Synthetic white latex																				
		Synthetic blue latex																				
		Synthetic black latex																				

REFERENCES

- Birger R.; D. Wout and A. El-Houssaine. 2013. "A Matheuristic for Aggregate Production-distribution Planning with Mould Sharing". *International Journal of Production Economics*, 145, Issue 1, 29-37.
- Dehau, X. and Y. Dar-Li. 2013. "Makespan Minization for Two Parallel Machines Model with a Periodic Availability Constraint : Mathematical Programming Model, Average-Case Analysis, and Anomalies". *International Journal of Applied Mathematical Modelling*, 7561-7567.
- Jirasukprasert, P.; J.A. Garza-Reyes; H. Soriano-Meier; H., and L. Rocha-Lona. 2012. "A Case Study of Defects Reduction in a Rubber Gloves Manufacturing Process by Applying Six Sigma Principles and DMAIC Problem Solving Methodology." *Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management*, 472-481.
- Haan, J. de.; G. De. Groot; E. Loo and M. Ypenburg. 2003. "Flows of Goods or Supply Chains; Lessons from the Natural Rubber Industry in Kerala, India." *International Journal of Production Economics*, Vol. 81-82, 185-194.
- Hsu,C,J.; W.H. Kuo and D.L. Yang. 2011. "Unrelated Parallel Machine Scheduling with Past-sequence-dependent Setup Time and Learning". *Applied Mathematical Modelling* 35, 1492-1496.
- Klomsae,S.; T. Somboonwiwat and W. Atthirawong. 2012. "Optimal multi-period Planning for Purchase and Storage of Rubber Latex with Perishability Constraints." *Proceedings of the 7th International Congress on Logistics and SCM Systems*, 244-250.
- Rattanapana, C.; T.T. Suksarojb, W.Ounsanehab. 2012. "Development of Eco-efficiency Indicators for Rubber Glove Product by Material Flow Analysis". *Procedia - Social and Behavioral Sciences* 40, 99 - 106.
- Wen-Chiung, L.; C. Mei-Chi and Y.Wei-Chang. 2012. "Uniform Parallel-machine Scheduling to Minimize Makespan with Position-based Learning Curves". *International Journal of Information Computer and Industrial Engineering*, 813-818.
- Yan, H.S.; H.X. Wang and X.D. Zhang. 2013. "Simultaneous Batch Splitting and Scheduling on Identical Parallel Production Lines". *Information Science* 221, 501-519.

AUTHOR BIOGRAPHIES



TUANJAI SOMBOONWIWAT is an associate Professor in the Industrial Management section, Department of Production Engineering, King Mongkut's University of Technology Thonburi, Thailand. She received her Ph.D. in Industrial Engineering from Corvallis, Oregon State University, USA. Her research interests include green supply chain and logistics, business process and applications of operations research. Her e-mail address is : tuanjai.som@kmutt.ac.th.



CHORKAEW JATURANONDA is a faculty member at Department of Production Engineering, King Mongkut's University of Technology Thonburi, Thailand. She received a bachelor's degree in Applied Mathematics from King Mongkut's University of Technology Ladkrabang (Thailand), a master's degree in Industrial Engineering from University of Texas at Arlington (USA), and a Ph.D. degree in industrial engineering from Sirindhorn International Institute of Technology, Thammasat University (Thailand). Her research interests include applied operations research, logistics management, and Decision Support Systems. Her e-mail address is : chorkaew.jat@kmutt.ac.th



KANOGKAN LEEROJANAPRAPA was born in Bangkok, Thailand and went to the Thammasat University, where she studied Applied Statistics and obtained her degree in 1999. She continued to study Master degree in Statistics in Chulalongkorn University and obtained her degree in 2002. She worked for four years for King Mongkut's Institute of Technology Ladkrabang (KMITL) before doing her PhD in 2008, University of Strathclyde, UK. After her graduation, she returned to KMITL where she is now a lecturer in Statistics department. Her e-mail address is : klkanogk@kmitl.ac.th



NATTAPONG CHOTPAN was a graduate student in Industrial and Manufacturing Systems Engineering program at the Department of Production Engineering, King Mongkut's University of Technology Thonburi, Thailand. He received his bachelor's degree in Industrial Engineering from Ramkhamhaeng University. His research interests are in supply chain management, production planning and optimization model. His e-mail address is : nattapong-ch@hotmail.com.

OPTIMAL SCHEDULING OF TWO-STAGE REENTRANT HYBRID FLOW SHOP FOR HEAT TREATMENT PROCESS

Noppachai Chalardkid
Tuanjai Somboonwiwat
Chareonchai Khompatraporn
Department of Production Engineering,
King Mongkut's University of Technology Thonburi (KMUTT), Bangkok 10140 Thailand
E-mail: tuanjai.som@kmutt.ac.th, charoenchai.kho@kmutt.ac.th

KEYWORDS

Optimal Scheduling, Reentrant Hybrid Flow Shop, Heat Treatment Process.

ABSTRACT

The reentrant hybrid flow shop for a heat treatment process is considered in this study. We consider job scheduling in a reentrant hybrid flow shop problem that consists of two stages in series. The first stage is washing, followed by heat treating in the second stage. Each job passes through the first and second stages, respectively, and then re-enters the first stage one more time. Since the first stage must process the jobs twice (with different processing times depending upon the type of the jobs), it becomes the bottleneck in this flow shop problem. To resolve this problem, the jobs needed to be better sequenced to balance the load among the first and the second stages. The objective is to minimize makespan of a set of jobs and increase the utilization of the both stages. This problem was formulated as a mixed integer program (MIP). The results from the data set show that the utilization of the second stage (heat treating) increased from 79.5% to their full capacity at 100%, exceeding the target set by the company at 95%.

INTRODUCTION

Heat treatment is a technique to enhance the properties of materials to a desired level. The heat treatment process is widely used with automotive parts to increase their strength. The process consists of a heating and cooling cycle. In practice, the parts are heated in heat furnaces and cooled by passing through washing machines. However, prior to the heating parts need to be washed to clean the parts. This production process is equivalent to a two-stage flow shop, whereas the first stage is washing, then heating, and re-washing (reentrance). Each stage may consist of several machines working in parallel. This problem is called a reentrant hybrid flow shop scheduling (RHFSS) (Choi et al. 2009).

The RHFSS problem was studied by Watanakich 2001. He studied a scheduling problem for a two-stage hybrid

flow shop with machine setup time, and solved the problem using a heuristic. The heuristic is composed of two phases. The first phase constructs a schedule and the second phase assigns jobs with setup time consideration according to the constructed schedule. Vignier et al. 1996 applied a branch and bound based algorithm to schedule jobs in multi-stages flow shop to minimize the makespan. Pan and Chen 2005 formulated a multi-stage reentrant hybrid flow shop as a mixed integer program (MIP). Their study also aims to minimize the makespan of a set of jobs. Choi et al. 2008 studied a mixed binary integer program of two-stage reentrant hybrid flow shop in which a due-date constraint was added to prevent tardiness. Their application was a coating process in automotive manufacturing that jobs were first coated in the paint shop, passed through an oven, and then coated the second time in the paint shop. Yalaoui et al. 2009 proposed several algorithms to solve a scheduling problem of a re-entrant production line. Their objective was to minimize the total tardiness. The performance of their heuristic was better than those of the EDD algorithm.

This paper presents a two-stage reentrant hybrid flow shop problem with due date constraints. In this flow shop all new jobs pass through the first and second stages. Then jobs reenter the first stage again. After the reentrance, jobs exit the flow shop. There are several identical machines (and furnaces) working in parallel at these stages. Jobs require different heat treatment times, but the time needed to wash them are the same. This problem is formulated as a mathematical program, and solved by LINGO commercial software.

The organization of this paper is as the following. The next section describes the two-stage re-entrant hybrid flow shop of the heat treatment process. Then the problem is formulated as a mixed integer program. A numerical example is presented to illustrate an application of the formulation. Finally, the paper is concluded.

PROBLEM DESCRIPTION

The heat treatment process is composed of three sequential processes. The first process is washing. Then the parts are heat treated, and passed through the second washing. The washing machines and heat treating furnaces were arranged as a two-stage hybrid flow shop. The flow of the jobs starts from entering through one of the washing machines (working in parallel) to clean the parts. Then each of the jobs will pass through one of the heat treating furnaces. Then each of the jobs re-enters any of the washing machines one more time, as shown in Figure 1.

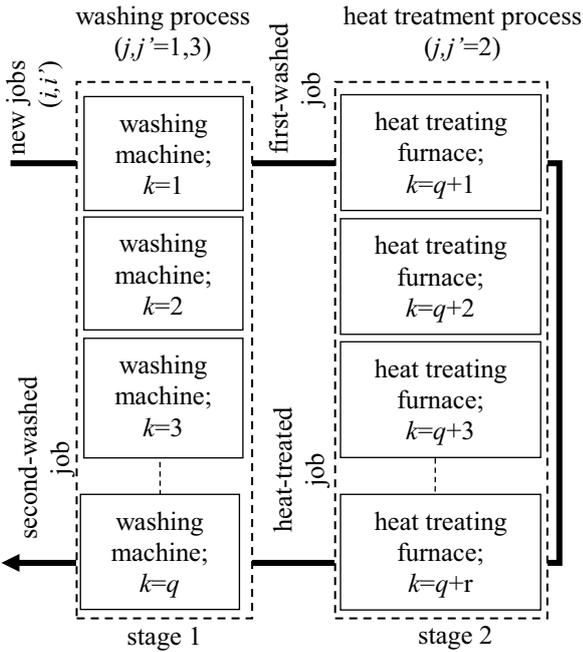


Figure 1: Two-stage Reentrant Hybrid Heat Treatment Flow Shop

The objective is to schedule jobs to these sets of the washing machines and the heat treatment furnaces so that the total makespan of all the jobs is minimized. This implies that the jobs must be sequenced to balance the utilization among these machines. To sequence the jobs, the starting times of each job to the washing machines and the furnaces, and their order must be determined. Since the jobs reenter the washing machines, each job needs two enter times to the washing machines. One for the first pass and the other for the second pass.

The next section describes this two-stage reentrant hybrid flow shop with various constraints as a mathematical program.

MATHEMATICAL FORMULATION

The mathematical model is a system of equations established to solve the problem of optimal scheduling of two-stage reentrant hybrid flow shop for heat treatment process. As earlier mentioned, the objective is to minimize the makespan.

Let X_{ijk} be the assignment of job i to machine (or furnace) k for processing operation j , and S_{ij} be the starting time of job i for processing operation j . The due date of job i is denoted by D_i . The followings are a complete list of all notations.

Indices

- i, i' jobs : $i, i' = 1, 2, 3, \dots, n$
- j, j' operations : $j, j' = 1$: first washing
 $j, j' = 2$: heat treatment
 $j, j' = 3$: second washing
- k machines at stage 1 : $k = 1, 2, 3, \dots, q$
furnaces at stage 2 : $k = q + 1, q + 2, q + 3, \dots, q + r$

Parameters

- D_i due-date of jobs i
- M very large positive number
- n number of jobs
- P_{ij} processing time of job i processed at operation j
- q number of machines at stage 1
- r number of furnaces at stage 2

Decision Variables

- S_{ij} starting time of job i processed at operation j
- $X_{ijk} = 1$ if job i processed at operation j on machine or furnace k , and 0 otherwise
- $Y_{ij'i'} = 1$ if job i processed at operation j precedes job i' that processed at operation j' , and 0 otherwise

Objective Function

The objective function is to minimize the makespan. The equation is :

$$\text{Minimize } C_{max} \quad (1)$$

where C_{max} is described by Equations (10) and (11)

Constraints

1. For each job, the starting time of the next operation j' must follow the finish time of the current operation j :

$$\sum_{k=1}^q X_{ijk} (S_{ij} + P_{ij}) \leq \sum_{k=q}^{q+r} X_{ij'k} S_{ij'}, \quad \forall i, j = 1 \quad \text{and } j' = 2 \quad (2)$$

$$\sum_{k=q}^{q+r} X_{ijk} (S_{ij} + P_{ij}) \leq \sum_{k=1}^q X_{ij'k} S_{ij'}, \quad \forall i, j = 2 \quad \text{and } j' = 3 \quad (3)$$

2. Each job must be processed at operation j by only one machine (or furnace):

$$\sum_{k=1}^q X_{ijk} = 1, \forall i \text{ and } j = 1, 3 \quad (4)$$

$$\sum_{k=q}^{q+r} X_{ijk} = 1, \forall i \text{ and } j = 2 \quad (5)$$

3. For the first stage, a job can be processed only on a machine at any time.

$$(2 - X_{ijk} - X_{i',j',k})M + (1 - Y_{ij'i'})M + S_{i'j'} - S_{ij} \leq P_{ij}, \text{ where} \\ 1 \leq i < i' \leq n, j, j' = 1, 3 \text{ and } k = 1, 2, 3, \dots, q \quad (6)$$

$$(2 - X_{ijk} - X_{i',j',k})M + Y_{ij'i'}M + S_{ij} - S_{i'j'} \leq P_{i'j'}, \text{ where} \\ 1 \leq i < i' \leq n, j, j' = 1, 3 \text{ and } k = 1, 2, 3, \dots, q \quad (7)$$

4. For the second stage, a job can be processed only on a machine at any time.

$$(2 - X_{ijk} - X_{i',j',k})M + (1 - Y_{ij'i'})M + S_{i'j'} - S_{ij} \leq P_{ij}, \text{ where} \\ 1 \leq i < i' \leq n, j = 2 \text{ and } k = q+1, q+2, q+3, \dots, q+r \quad (8)$$

$$(2 - X_{ijk} - X_{i',j',k})M + Y_{ij'i'}M + S_{ij} - S_{i'j'} \leq P_{i'j'}, \text{ where} \\ 1 \leq i < i' \leq n, j = 2 \text{ and } k = q+1, q+2, q+3, \dots, q+r \quad (9)$$

5. For each job, the completion time must be less than the makespan.

$$\sum_{k=1}^q X_{ijk} (S_{ij} + P_{ij}) \leq C_{max}, \forall i \text{ and } j = 3 \quad (10)$$

$$C_{max} \geq 0 \quad (11)$$

6. For each job, the completion time must not exceed the due date.

$$\sum_{k=1}^q X_{ijk} (S_{ij} + P_{ij}) \leq D_i, \quad \forall i \text{ and } j = 3 \quad (12)$$

7. Constraints describing the decision variables.

$$S_{ij} \geq 0, \forall i, j \quad (13)$$

$$X_{ijk} \in \{0, 1\}, \forall i, j = 1, 3 \text{ and } k = 1, 2, 3, \dots, q \quad (14)$$

$$X_{ijk} \in \{0, 1\}, \forall i, j = 2 \text{ and } k = q+1, q+2, q+3, \dots, q+r \quad (15)$$

$$Y_{ij'i'} \in \{0, 1\}, 1 \leq i < i' \leq n \text{ and } j, j' = 1, 3 \quad (16)$$

$$Y_{ij'i'} \in \{0, 1\}, 1 \leq i < i' \leq n \text{ and } j = 2 \quad (17)$$

NUMERICAL EXAMPLE

In this example, 15 jobs were processed on two washing machines and two heat treating furnaces. We denote $k = 1$ and 2 to represent the two washing machines, and $k = 3$ and 4 to represent the two heat treating furnaces. There are four different types of jobs depend on the heat treatment time; 180, 300, 420 and 600 minutes. Each job was washed for 45 minutes during the first washing and then another 45 minutes for the second washing. All jobs must be completed in four days (or 5,760 minutes)

as shown in the Table 1. We assume that the first job starts at time zero.

Table 1: Job Processing Times and Due Dates

Job (i)	Processing time (P _{ij}), min.			Due date (D _i), min.
	First washing j=1	Heat treatment j=2	Second washing j=3	
1	45	180	45	5,760
2	45	300	45	5,760
3	45	300	45	5,760
4	45	420	45	5,760
5	45	600	45	5,760
6	45	180	45	5,760
7	45	300	45	5,760
8	45	300	45	5,760
9	45	420	45	5,760
10	45	600	45	5,760
11	45	180	45	5,760
12	45	300	45	5,760
13	45	300	45	5,760
14	45	420	45	5,760
15	45	600	45	5,760

Results

LINGO software is used to solved this problem. The results are shown in Table 2. The table shows the starting time and completion time of the jobs, or their schedule. The makespan is then calculated by taking the difference of the starting time of the first job that enters the flow shop and the completion time of the last job that leaves the flow shop.

The results show the optimal schedule of the jobs in this two-stage reentrant hybrid flow shop by minimizing the makespan under a specific due date. Each job is assigned to a specific machine and a furnace every time it passes through these stages, and only allowed to access the machine or the furnace at a specific starting time. For example Job 1 enters the washing machine No.2 ($k=2$) for 45 minutes, starting at the time 1,500 minute, then passes through the heat treating furnace No.2 ($k=4$) for 180 minutes immediately after leaving the washing machine. The job leaves the furnace at time 1,725 minute. Then the job waits to reenter the washing process through the first washing machine ($k=1$) at time 2,700 minute. The job is completed at time 2,745 minute. Other jobs can be described similarly.

From the table, the first jobs that enter the flow shop are Jobs 10 and 12, and the last jobs that leave are Jobs 2 and 4. The makespan of all the jobs is 2,790 minutes. With the job schedule details in Table 2, the utilization of heat treating furnaces is at 100% and all the jobs are completed prior to the due date. Details of the schedule of all the jobs by washing machines and heat treating furnaces can be found in Figure 2.

Table 2: Job Schedule

Job (i)	Process (j)	Processing time (P_{ij}), min.	Job assignment, X_{ijk}				Starting time (S_{ij}), min	Completion time, min
			Washing machine		Heat treating furnace			
			k=1	k=2	k=3	k=4		
1	1	45	0	1	0	0	1,500	1,545
	2	180	0	0	0	1	1,545	1,725
	3	45	1	0	0	0	2,700	2,745
2	1	45	0	1	0	0	2325	2,370
	2	300	0	0	1	0	2,445	2,745
	3	45	1	0	0	0	2,745	2,790
3	1	45	1	0	0	0	2,100	2,145
	2	300	0	0	1	0	2,145	2,445
	3	45	0	1	0	0	2,445	2,490
4	1	45	0	1	0	0	2,280	2,325
	2	420	0	0	0	1	2,325	2,745
	3	45	0	1	0	0	2,745	2,790
5	1	45	1	0	0	0	480	525
	2	600	0	0	1	0	525	1,125
	3	45	0	1	0	0	2,370	2,415
6	1	45	0	1	0	0	300	345
	2	180	0	0	1	0	345	525
	3	45	0	1	0	0	2,235	2,280
7	1	45	1	0	0	0	1,170	1,215
	2	300	0	0	1	0	1,845	2,145
	3	45	0	1	0	0	2,145	2,190
8	1	45	0	1	0	0	1,080	1,125
	2	300	0	0	1	0	1,125	1,425
	3	45	0	1	0	0	1,425	1,470
9	1	45	0	1	0	0	90	135
	2	420	0	0	1	0	1,425	1,845
	3	45	1	0	0	0	1,845	1,890
10	1	45	1	0	0	0	0	45
	2	600	0	0	0	1	45	645
	3	45	0	1	0	0	645	690
11	1	45	1	0	0	0	90	135
	2	180	0	0	0	1	645	825
	3	45	1	0	0	0	2,655	2,700
12	1	45	0	1	0	0	0	45
	2	300	0	0	1	0	45	345
	3	45	1	0	0	0	2,610	2,655
13	1	45	1	0	0	0	135	180
	2	300	0	0	0	1	825	1,125
	3	45	1	0	0	0	1,125	1,170
14	1	45	1	0	0	0	45	90
	2	420	0	0	0	1	1,125	1,545
	3	45	1	0	0	0	1,545	1,590
15	1	45	0	1	0	0	45	90
	2	600	0	0	0	1	1,725	2,325
	3	45	1	0	0	0	2,325	2,370

Although there are merely 15 jobs in this example, the problem requires approximately two and a half hours to obtain the final optimal solution by LINGO. First, the software found an upper bound of the optimal solution, then this bound was entered to the software to find the final solution. This shows that scheduling jobs in this type of flow shop is not an easy task.

CONCLUSION

A two-stage reentrant hybrid flow shop with washing and heat treatment processes as the two stages is addressed in this paper. The reentrance occurs only at the first stage, after the jobs leave the second stage. The problem is formulated as a mixed integer program. A numerical example is used to illustrate the application of the formulation, and solved by a commercial software, LINGO. In the example, 15 jobs are scheduled to minimize the makespan under a specific due date. It was found that the schedule of the jobs to the first washing affects the utilization of the heat treating furnaces, whereas the schedule of the second washing is critical to the final makespan. The schedule obtained balances the workload among the machines and furnaces. As a consequence, the utilization of the furnaces increased from 79.5% to 100% (theoretically), exceeding the 95% target set by the company. This model can also be extended to a flow shop with more than two stages, and the jobs may reenter to any of the stages. For large-scale problems, a heuristic may be developed to quickly solve the problem because a multi-stage reentrant hybrid flow shop scheduling is considered as a hard problem.

REFERENCES

Choi, H.S., Kim, H.W., Lee, D.H., Yoon, J., Yun, C.Y. and K.B. Chae. 2008. "Scheduling algorithms for two-stage reentrant hybrid flow shops: minimizing makespan under the maximum allowable due dates," *International Journal Advanced Manufacturing Technology*, Vol.42, 963–973.

Pan, J.C. and J. Chen. 2005. "Mixed binary integer programming formulations for the reentrant job shop scheduling problem," *Computers & Operations Research*, Vol.32, 1197–1212.

Vignier, A., Dardilhac, D., Dezalay, D. and C. Proust. 1996. "A branch and bound approach to minimize the total completion time in a k-stage hybrid flowshop," *Proceedings of 1996 IEEE Conference on Emerging Technologies and Factory Automation*, Vol.1, pp. 215-220.

Watanakich, P., 2001, Scheduling for a Two-Stage Hybrid Flow Shop With Machine Setup Time, Master of Engineering Thesis, Industrial Engineering, Faculty of Engineering, Kasetsart University, pp. 1-55.

Yalaoui, N., Camara, M., Amodeo, L., Yalaoui, F. and H. Mahdi. 2009, "New heuristic for scheduling re-entrant production lines," *Proceedings of the International Conference on Computers & Industrial Engineering (CIE 2009)*, pp.199-204.

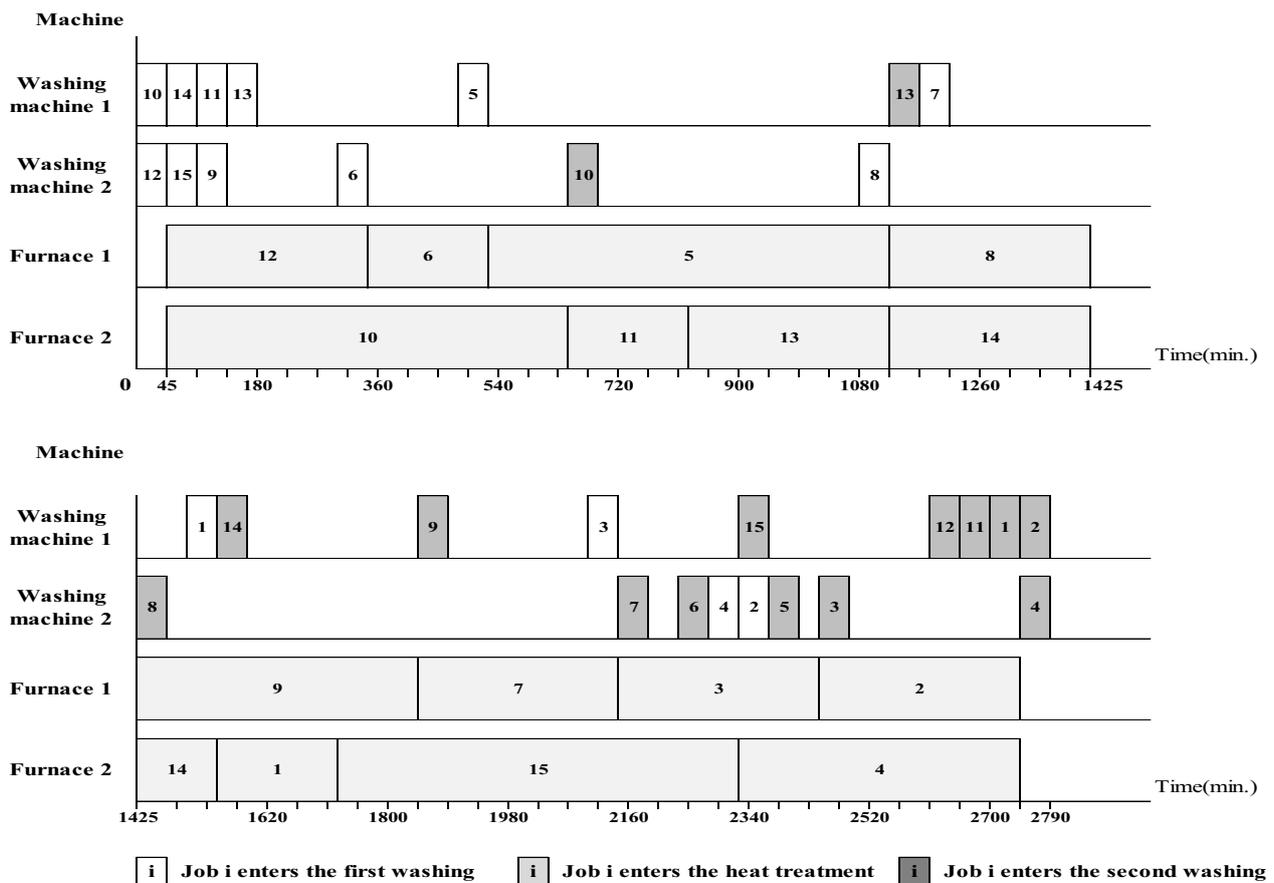


Figure 2: Job Schedule by Machines and Furnaces

AUTHOR BIOGRAPHIES



NOPPACHAI CHALADKID was a graduate student in Industrial and Manufacturing Systems Engineering program at the Department of Production Engineering, King Mongkut's University of Technology Thonburi, Thailand. He received his bachelor's degree in Electronics and Telecommunication Engineering from the same university. His research interests are in supply chain management and applied statistics. His e-mail address is: ene46221226@yahoo.com.



TUANJAI SOMBOONWIWAT is an Associate Professor in the Industrial Management section, Department of Production Engineering Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand. She received her M.Eng. in Industrial Engineering from Chulalongkhorn University, Thailand and Ph.D. in Industrial Engineering from Corvallis, Oregon State University, USA. Her research interests include green supply chain and logistics, business process and

applications of operations research. Her e-mail address is: tuanjai.som@kmutt.ac.th.



CHAROENCHAI KHOMPATRAPORN holds a Ph.D. from University of Washington, USA. His research interests include supply chain and logistics management, applied operations research, optimization algorithms, and industrial sustainable operations management. He has been working closely with both public and private sectors such as the hard disk drive industry, the automotive industry, and Thai Red Cross. His e-mail address is: charoENCHAI.kho@kmutt.ac.th

MATHEMATICALLY MODELLING HCG IN WOMEN WITH GESTATIONAL TROPHOBLASTIC DISEASE USING EXPONENTIAL INTERPOLATION

Catherine Costigan
Sabin Tabirca
School of Mathematics
University College Cork

John Coulter
Cork University Hospital
University College Cork

Ernest Scheiber
Department of Informatics
Transylvania University of Brasov
Romania

KEYWORDS

Mathematical Modelling; Gestational Trophoblastic Disease; Logarithmic Transformation; hCG; Exponential Interpolation

ABSTRACT

Exponential interpolation is a problem in mathematics which can have very useful results in many areas. One area which is particularly of interest here is using exponential interpolation to model human chorionic gonadotropin (hCG) levels in women with gestational trophoblastic disease as it has been previously reported that the hCG measurements in these women follow an exponential curve.

INTRODUCTION

Interpolation is the process of estimating the value of a function, $f(x)$, whose value is known at specific points, say $x_0 < x_1 < \dots < x_n$, by a second function $g(x)$ so that $f(x_i) = g(x_i)$ for each $i \in \{0, 1, \dots, n\}$. The values of the function $f(x)$ at other points in the interval (x_0, x_n) can then be estimated using the function $g(x)$. In most interpolation schemes, the function $g(x)$ is found to be unique and either an equation or an algorithm can be used to calculate it. For example, the linear, polynomial and spline interpolation methods provide nice equations for the unique function $g(x)$. Furthermore, theoretical results are also available for these methods to evaluate the estimation error of $\|f(x) - g(x)\|$ (see (Phillips, 2003) for a good review of interpolation methods).

Interpolation is used in many different areas including economics, engineering and medical simulation for a variety of reasons. One reason that interpolation is used is if the actual function $f(x)$ is known explicitly but is too complicated to use in a computer programme. In this case an interpolating function $g(x)$ that agrees closely with the actual function is used. Another situation where interpolation is used is when the value of the function $f(x)$ is only known at certain points. In this case an interpolating

function $g(x)$ is used so that the function can be estimated at other points (Szidarovszky & Yakowitz, 2013).

Exponential interpolation considers the function $g(x)$ as a linear combination of exponentials

$$g(x) = B + \sum_{i=1}^n A_i \cdot e^{\alpha_i \cdot x}. \quad (1)$$

In this case, the *Exponential Interpolation* problem can be formulated as follows:

Definition 1. Given a function $f : [a, b] \rightarrow \mathbb{R}$ which is twice differentiable and $2n + 1$ points

$$a \leq x_0 < x_1 < \dots < x_{2n+1} \leq b$$

then find the function $g(x)$ given by equation 1 so that $f(x_i) = g(x_i)$, $i = 1, 2, \dots, 2n + 1$.

Usually, in exponential interpolation, it is required to find a method to calculate the $2n + 1$ coefficients, B and $A_i, \alpha_i, i = 1, 2, \dots, 2n$, and then to prove that these coefficients are unique. It is also useful to have an evaluation for the interpolation error $|f(x) - g(x)|$.

Exponential Interpolation, often called interpolation with Dirichlet polynomials, has been investigated quite intensively for the last decades in order to solve the above problems (Ammar, Dayawansa, & Martin, 1991), however the interpolation methods proposed are either difficult to implement or computationally expensive. More effective approaches can be developed by some direct approaches especially in some particular cases of equation 1.

One particular form is given by $g(x) = A \cdot e^{-\alpha \cdot x} + B$, which occurs in various technical and medical decay problems. A problem of interest for us is the variation of the hCG marker for Gestational Trophoblastic Disease (GTD). GTD is a term used to describe a range of illnesses in women, such as complete and partial hydatidiform mole, choriocarcinoma, placental site trophoblastic tumour and epithelioid trophoblast tumour. All of these conditions arise from human trophoblastic tissue (Young et al., 2011). Upon diagnosis, the hydatidiform mole

should be surgically evacuated. Usually, after this procedure the hCG levels will drop exponentially (Schoeberl, 2007). Interpreting the drop in hCG levels is a useful way to decide on a treatment plan for the patients, and to decide if the use of chemotherapy is required (Schoeberl, 2007).

The aim of this study is to mathematically model hCG level in women with gestational trophoblastic disease using Exponential Interpolation. It has been suggested previously in (Almufti et al., 2014) that in the future, mathematical models may be used to describe tumour biomarker production. This could be in relation to tumour size, the effectiveness of treatment and biomarker elimination. It was even thought that based on a few time points for each patient, mathematical models may enable early prediction of changes in disease burden and treatment efficacy, however little concrete research work has been done in this direction. It was reported in (You et al., 2010) and (You et al., 2013) that the model that hCG follows after evacuation is of the form

$$g(t) = Ae^{-\alpha t} + B. \quad (2)$$

Firstly, this article investigates the exponential interpolation based on equation 2 in order to calculate the coefficients A, B and α . Then we will prove that these coefficients are unique and will evaluate the interpolation error. Finally, we present how this interpolation can be applied to predict the evolution of the hCG marker in GTD.

EXPONENTIAL INTERPOLATION FOR $n = 1$

The interpolation process needs to find the exponential function

$$g(t) = Ae^{-\alpha t} + B \quad (3)$$

that goes through three points (x_i, y_i) e.g. $g(x_i) = y_i$, $i = 0, 1, 2$. We assume that

$$x_0 < x_1 < x_2 \Rightarrow y_0 > y_1 > y_2. \quad (4)$$

Proposition 1. *Given the points $(x_0, y_0), (x_1, y_1)$ and (x_2, y_2) , the parameters A and B can be calculated using the formulas so that $y_i = g(x_i)$ for each $i \in \{0, 1, 2\}$.*

$$A = \frac{y_1 - y_0}{e^{-\alpha x_1} - e^{-\alpha x_0}}, \quad B = y_0 - Ae^{-\alpha x_0} \quad (5)$$

and α can be found by finding a root of the equation

$$\frac{e^{-\alpha(x_1-x_0)} - 1}{e^{-\alpha(x_2-x_0)} - 1} = \frac{y_1 - y_0}{y_2 - y_1} \quad (6)$$

Proof. Insisting the function passes through the points $\{x_0, y_0\}, \{x_1, y_1\}$ and $\{x_2, y_2\}$ means the equations

$$\begin{aligned} Ae^{-\alpha x_0} + B &= y_0 \\ Ae^{-\alpha x_1} + B &= y_1 \\ Ae^{-\alpha x_2} + B &= y_2 \end{aligned} \quad (7)$$

must be satisfied. Solving for B gives $B = y_0 - Ae^{-\alpha x_0}$ and substituting it into the second equation to obtain $A = \frac{y_1 - y_0}{e^{-\alpha x_1} - e^{-\alpha x_0}}$. To get the equation for α both of these can be substituted into the third equation to give

$$\frac{y_1 - y_0}{e^{-\alpha x_1} - e^{-\alpha x_0}} e^{-\alpha x_2} + y_0 - \frac{y_1 - y_0}{e^{-\alpha x_1} - e^{-\alpha x_0}} e^{-\alpha x_0} = y_2.$$

Tidying this up gives

$$\frac{e^{-\alpha(x_1-x_0)} - 1}{e^{-\alpha(x_2-x_0)} - 1} = \frac{y_1 - y_0}{y_2 - y_0} \quad (8)$$

as desired. \square

Theorem 1. *Equation 6 has a unique solution for α .*

Proof. From equation 2 $g(x) = Ae^{-\alpha x} + B$. Then

$$g'(x) = -A\alpha e^{-\alpha x} \quad (9)$$

From this it is clear that as long as $A \cdot \alpha \neq 0$ then $g(x)$ is strictly monotone. Since for $x_0 < x_1 < x_2 \Rightarrow y_0 > y_1 > y_2$ we have that $g(x)$ is decreasing i.e. $A \cdot \alpha > 0$. Concentrating on this case, let

$$p = x_1 - x_0 > 0 \quad (10)$$

$$q = x_2 - x_0 > 0, q > p \quad (11)$$

$$k = \frac{y_1 - y_0}{y_2 - y_0} \in (0, 1) \quad (12)$$

Now, equation 6 can be rewritten in terms of p, q and k to give

$$m(\alpha) = \frac{e^{-p\alpha} - 1}{e^{-q\alpha} - 1} - k = 0. \quad (13)$$

The function $m(\alpha)$ can be extended at 0 by $m(0) = \frac{p}{q} - k$ so that it is continuous and differentiable over \mathbb{R} . By looking at the behaviour of $m(\alpha)$ it can be shown that a unique solution for $m(\alpha) = 0$ exists.

$$\lim_{\alpha \rightarrow \infty} m(\alpha) = 1 - k > 0 \quad (14)$$

and

$$\lim_{\alpha \rightarrow -\infty} m(\alpha) = -k < 0. \quad (15)$$

As a consequence of this, a solution of equation 8 can be found. In order to show that the solution for α is unique it will be shown that the equation $m(\alpha)$ is strictly monotone. Finding the derivative of $m(\alpha)$ gives

$$m'(\alpha) = e^{(q-p)\alpha} \cdot \frac{e^{p\alpha} - 1}{e^{q\alpha} - 1} \cdot \left(\frac{p}{e^{p\alpha} - 1} - \frac{q}{e^{q\alpha} - 1} \right). \quad (16)$$

Multiplying the top and bottom by α gives

$$m'(\alpha) = e^{(q-p)\alpha} \cdot \frac{e^{p\alpha} - 1}{e^{q\alpha} - 1} \cdot \left(\frac{p\alpha}{e^{p\alpha} - 1} - \frac{q\alpha}{e^{q\alpha} - 1} \right) \cdot \frac{1}{\alpha}. \quad (17)$$

Examining each part of $m'(\alpha)$ to check to see if $m'(\alpha)$ is positive or negative, since $q > p$ it is clear that $e^{(q-p)\alpha}$ and $\frac{e^{p\alpha} - 1}{e^{q\alpha} - 1}$ are always positive. $\frac{1}{\alpha}$ is positive for positive

α and negative for negative α so it just remains to find the sign of $\frac{p\alpha}{e^{p\alpha}-1} - \frac{q\alpha}{e^{q\alpha}-1}$.

Looking at the function $\psi(t) = \frac{t}{e^t-1}$, $\psi'(t) = \frac{1-e^t(t-1)}{(e^t-1)^2}$, we can find that $\psi'(t)$ is always negative as the numerator is negative and the denominator is positive. This implies that the function $\psi'(t)$ is decreasing. This means that $\frac{p\alpha}{e^{p\alpha}-1} - \frac{q\alpha}{e^{q\alpha}-1}$ is positive for positive α and negative for negative α .

So overall $m'(\alpha) > 0$ for all values of α , which means that $m(x)$ is strictly increasing. So $m(\alpha)$ is injective so there is a unique solution to equation 6. \square

Theorem 2. Given a function $f : [a, b] \rightarrow \mathbb{R}$ n times differentiable and n points $\{x_i\}_{i=0}^{n-1}$ so that $f(x_i) = 0$ for all $i \in \{0, \dots, n-1\}$, then for all $x \in (a, b)$ there is a number $c \in (a, b)$ so that

$$f(x) = \frac{f^{(n)}(c)}{n!} \prod_{i=0}^{n-1} (x - x_i) \quad (18)$$

This is just a consequence of the interpolation error theorem (Phillips, 2003) in which the interpolation polynomial is 0 since all the values $f(x_i) = 0$ are 0.

Proposition 2. Given a function $f : [a, b] \rightarrow \mathbb{R}$ 3 times differentiable and $g(x)$ the exponential function defined in equation 2 so that $f(x_i) = g(x_i) = y_i$, $i = 0, 1, 2$ (with A, B and α satisfying equations 7). Then the interpolation error is provided by:

$$\frac{|f(x) - g(x)|}{\max_{x \in (a, b)} \frac{|f'''(x) + A\alpha^3 e^{-\alpha x}|}{6} |(x-x_1)(x-x_2)(x-x_3)|} \quad (19)$$

Proof. Define the function

$$m(x) = f(x) - g(x) \quad (20)$$

with the properties $m(x)$ is 3 times differentiable and $m(x_0) = m(x_1) = m(x_2) = 0$. By theorem 2 there is some $c \in (x_0, x_2)$ so that

$$m(x) = \frac{m'''(c)}{6} (x-x_0)(x-x_1)(x-x_2) \quad (21)$$

This means that

$$\frac{|f(x) - g(x)|}{\max_{x \in (x_0, x_2)} \frac{|m'''(x)|}{6} |(x-x_0)(x-x_1)(x-x_2)|} \quad (22)$$

\square

Remark 1. As a consequence of proposition 2

$$\begin{aligned} & |f(x) - m(x)| < \\ & \frac{\max_{x \in (x_0, x_2)} |f'''(x)| + A\alpha^3 e^{-\alpha x_0}}{6} |(x-x_0)(x-x_1)(x-x_2)| \\ & \leq \frac{\max_{x \in (x_0, x_2)} |f'''(x)| + A\alpha^3}{6} |(x-x_0)(x-x_1)(x-x_2)|. \end{aligned} \quad (23)$$

From experiments it was found that the value of A is very large, at least in the order of 10^3 . So this error term is not small enough to ensure that the errors would be sufficiently small when using this method.

IMPLEMENTING EXPONENTIAL INTERPOLATION FOR $n = 1$

The method described in proposition 1 was implemented in C# in order to fit the model 2 to the data. The Newton-Raphson method was used on equation 6 in order to find the root. This method is an iterative root finding technique that uses an initial guess together with the value of the function at that point and its derivative at the point in order to make a better approximation to the root. The $n + 1^{th}$ guess is found using the formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (24)$$

The process continues until the difference $x_{n+1} - x_n < \epsilon$, where ϵ is the accuracy of the root desired. On implementing the Newton-Raphson method, the values of the derivatives of 6 were estimated numerically at each iteration (Szidarovszky & Yakowitz, 2013) (Press, Teukolsky, Vetterling, & Flannery, 1992).

Algorithm 1 Finding A, B and α using Interpolation

input: $x_0, y_0, x_1, y_1, x_2, y_2$
 $\alpha = \text{newtonraphson}(\text{equation } 8)$
 $A = \text{findA}(\alpha)$ {using equation 5}
 $B = \text{findB}(\alpha, A)$ {using equation 5}
output: A, B, α

In order to test if the method of interpolation works well for fitting a curve to data points some points were chosen from a curve in the form of 2 with known values for A, B and α . In this case, the values for A, B and α chosen were 1000, 7 and 0.5 respectively. Using the method described in algorithm 1 the parameters

$$\begin{aligned} A &= 999.999 \\ B &= 7.0009 \\ \alpha &= 0.5000 \end{aligned}$$

were found. This shows that this method for interpolation works well. The error in the estimated parameters is negligible.

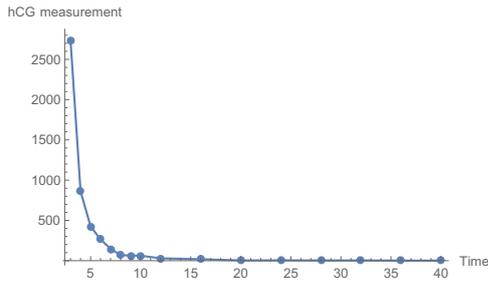


Figure 1: hCG measurements for a Typical Patient

MODELLING hCG MARKER FOR GTD

Upon diagnosis, the hydatidiform mole should be surgically evacuated. Usually, after this procedure the hCG levels will drop exponentially (Schoeberl, 2007). Sometimes, however, after a possible drop at the start the hCG levels will plateau or rise again. Sustained high levels of hCG are an identifier of gestational trophoblastic neoplasia (GTN), these high levels of hCG have also been an indicator for the need for chemotherapy (Schoeberl, 2007). Interpreting the drop in hCG levels is a useful way to decide on a treatment plan for the patients, and to decide if the use of chemotherapy is required (Schoeberl, 2007).

Each woman in the UK, and now in Ireland, diagnosed with GTD should be registered with a specialist centre for hCG surveillance as it is necessary to monitor hCG levels in these women since it is known that about 15% of women with a complete hydatidiform mole will require chemotherapy (Seckl, Sebire, & Berkowitz, 2010). Currently in these centres in the UK hCG is measured every 2 weeks following evacuation until the values are within normal range. After this, hCG concentrations are measured monthly for 6 months from the evacuation date (Seckl et al., 2010). Therefore, a number of blood tests (usually about 10 - 14) are conducted fortnightly in order to assess the evolution of the hCG marker. In most patients' cases, these hCG levels are decreasing suggesting that the chemotherapy treatment is not needed. In practice it often happens that a patient would miss some blood tests so that there may be a discontinuity of the data. In this case, we can apply exponential interpolation to best fit the curve $g(x) = Ae^{-\alpha x} + B$ through the available data, in order to estimate the missing data and to predict the future data.

A TYPICAL PATIENT

The hCG measurements for a typical patient are shown in figure 1.

Assume the patient has missed her blood test at week 5, but the results are known for week 3, 5 and 6. Using interpolation to find the coefficients A , B and α in the

model 2, the values found are

$$\begin{aligned} A &= 152,149.769 \\ B &= 222.544 \\ \alpha &= 1.368. \end{aligned}$$

Using these values in the model 2 it can be shown that this corresponds to a value of 385.36 at week 5, which means the relative error is 0.09. This is a good approximation to the actual value.

Another situation in which interpolation might be used is in predicting future measurements. Say the hCG measurements for weeks 3, 4 and 5 are known and an approximation for the measurement for week 6 is required. In this case the coefficients found are

$$\begin{aligned} A &= 193,159.163 \\ B &= 293.621 \\ \alpha &= 1.457. \end{aligned}$$

The actual hCG level for week 6 is 264, using this technique the approximation is 324.45, which corresponds to a relative error of 0.22. This is larger than the error for the previous test but this is to be expected as interpolation works best for approximating values within the interval used to generate the curve.

TESTING ON MULTIPLE WOMEN

Firstly, the second, third and fifth points from each data set were used for interpolation and then the fourth point was used to check the size of the error, the same way as in the previous section. The results of this can be seen in table 1. Each row in the table represents a different woman's data. The actual value is the fourth measured value of hCG in the blood of each patient. The error represents the difference between the actual fourth measurement and the value predicted using the curve that uses the values of A , B and α found using interpolation for each data set. The percentage error is the absolute value of the error, divided by the actual measurement. It can be seen that 7 out of the 21 points have a percentage error of less than 0.15. This seems to be reasonable, however, some of the points have very large errors, with 5 of the data points having percentage errors of more than 1. These errors are most likely due to the fact that the hCG levels drop exponentially fast so measurements taken in later weeks are small. This means that even a small deviation from the curve will result in a large relative error.

A test was then carried out using the second, third and fourth data points to find the values of the coefficients for the model 2, and then estimating the value for the fifth measurement. The results are shown in table 2.

data set	actual value	error	relative error
1	426	-40.64	0.09
2	1,250	-97.73	0.07
3	236	68.08	0.28
4	56	-11.91	0.21
5	5	-0.62	0.12
6	163	-21.32	0.13
7	295	-36.79	0.12
8	1,979	-770.59	0.38
9	388	1,003.83	2.58
10	34	-4.96	0.14

Table 1: Errors for predicting the fourth hCG measurement using interpolation

data set	actual value	error	relative error
1	264	60.45	0.22
2	357	158.42	0.44
3	290	-82.11	0.28
4	17	14.98	0.88
5	3	1.2	0.4
6	63	28.81	0.45
7	213	45.65	0.21
8	504	840.41	1.66
9	113	-113	1
10	26	5.312	0.20

Table 2: Errors for predicting the fifth hCG measurement using interpolation

Some women had measurement of 0 for their fifth measurement. These women were left out of this test. The results of this experiment are not quite as good as the previous experiment. This is to be expected however. Even though a lot of the cases have large errors, some of the cases have small enough errors, for instance case 14 has a relative error of just 0.2.

CONCLUSION

In this study interpolation was implemented. Some nice results were found about using interpolation to calculate values of A , B and α for the model 2 including a proof that a unique solution for the values of A , B and α exists.

The method described in this study was tested on simulated data and it works very well for finding the values of A , B and α . Estimating missing hCG measurements in the data was quite successful, with relative errors as small as 0.07 in the data sets tested. This method was not as accurate for predicting measurements outside of the values used to calculate the coefficients, In saying that, relative errors as small as 0.2 were observed.

There are two reasons the relative errors for predicting future measurements are larger than the relative errors for estimating missing measurements. The first is

that interpolation is known to be more accurate when estimating values within the interval used to calculate the coefficients. The other is that since the hCG levels are smaller for later weeks, the relative errors are larger for these values even if the absolute error stays the same.

References

- Almufti, R., Wilbaux, M., Oza, A., Henin, E., Freyer, G., Tod, M., ... You, B. (2014). A critical review of the analytical approaches for circulating tumor biomarker kinetics during treatment. *Annals of oncology*, 25(1), 41–56.
- Ammar, G., Dayawansa, W., & Martin, C. (1991). Exponential interpolation: theory and numerical algorithms. *Applied Mathematics and Computation*, 41(3), 189–232.
- Phillips, G. M. (2003). *Interpolation and approximation by polynomials* (Vol. 14). Springer Science & Business Media.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes: The art of scientific computing* (Cambridge). Cambridge Univ. Press.
- Schoeberl, M. R. (2007). A model for the behavior of β -hcg after evacuation of hydatidiform moles. *Gynecologic oncology*, 105(3), 776–779.
- Seckl, M. J., Sebire, N. J., & Berkowitz, R. S. (2010). Gestational trophoblastic disease. *The Lancet*, 376(9742), 717–729.
- Szidarovszky, F., & Yakowitz, S. J. (2013). *Principles and procedures of numerical analysis* (Vol. 14). Springer.
- You, B., Harvey, R., Henin, E., Mitchell, H., Golfier, F., Savage, P., ... Seckl, M. (2013). Early prediction of treatment resistance in low-risk gestational trophoblastic neoplasia using population kinetic modelling of hcg measurements. *British journal of cancer*, 108(9), 1810–1816.
- You, B., Pollet-Villard, M., Fronton, L., Labrousse, C., Schott, A.-M., Hajri, T., ... others (2010). Predictive values of hcg clearance for risk of methotrexate resistance in low-risk gestational trophoblastic neoplasias. *Annals of oncology*, 21(8), 1643–1650.
- Young, T., Coleman, R., Hancock, B., Drew, D., Wilson, P., Tidy, J., et al. (2011). Predicting gestational trophoblastic neoplasia (gtn): is urine hcg the answer? *Gynecologic oncology*, 122(3), 595–599.

Simulators for Virtual Prototyping and Training

vMannequin: a Fashion Store Concept Design Tool

Paolo Cremonesi, Franca Garzotto,
Marco Gribaudo, Pietro Piazzolla
Dipartimento di Elettronica,
Informazione e Bioingegneria
Politecnico di Milano
via Ponzio 31/32, Milano, Italy
Email: {paolo.cremonesi, franca.garzotto,
marco.gribaudo, pietro.piazzolla}@polimi.it

Mauro Iacono
Dipartimento di Scienze Politiche
Seconda Università degli Studi di Napoli
viale Ellittico 31, Caserta, Italy
Email: mauro.iacono@unina2.it

KEYWORDS

Concept Design; End User Development; 3D Computer Graphics

ABSTRACT

The fashion industry is one of the most flourishing fields for visual applications of IT. Due to the importance of the concept of look in fashion, the most advanced applications of computer graphics and sensing may fruitfully be exploited. The existence of low cost solutions in similar fields, such as the ones that empower the domestic video games market, suggest that analogous low cost solutions are viable and can foster innovation even in small and medium enterprises. In this paper the current state of development of vMannequin, a dynamic, user mimicking, user enacted virtual mannequin software solution, is presented. In order to allow users designing dress concepts, the application simulate the creation and fitting of clothes on virtual models. The interaction is sensor based, in order to both simplify the user interface, and create a richer involvement inside the application.

INTRODUCTION

The fashion industry is one of the application fields in which proper IT applications may be a strong enabling factor for innovation. Despite the information content of its final products is low, due to the peculiarly physical nature of pieces of cloths, there is a significant margin to be exploited by means of IT solutions in the production, sales, after sales support and services areas, due to the high information content of the related processes.

The availability of low cost computers capable of complex, real time graphical manipulation of realistic, dynamic 3D models, together with the availability of low cost and low impact positioning and movement sensors enables a number of innovative applications to support the processes of fashion design, being it in the phase of conceptual shaping of haute couture or mass market pieces of clothes or in the phase of personal outfit style shaping, and sales, in presence or by means of e-commerce web stores or virtual reality applications. Such solutions potentially offer many advantages; they may reduce the development cost for products, as much as the

piece of cloth is made of expensive materials or needs a complex manufacture, by reducing the need for prototyping; they may empower low cost tailor made production; they may meet the needs of customers with special needs by custom, unconventional, or even out of market pieces of clothes; they may allow small producers to emerge and acquire vertical market shares or reach a wider customer base; they may lower the barriers of the market for emerging stylists or firms; they can enable the creation virtual firms producing or assembling pieces of clothes or total looks by means of remote collaborative crowd production or crowd design. The benefits for a scattered (from the point of view of the size of the producers and from the point of view of price categories) market, like the Italian one is, are unpredictable, but possible developments may reasonably be considered to be interesting.

The availability of visual tools that allow the simulation of dressing up a customized mannequin and the effects of fabric, colour, cut on it while its movements mimic the movements of the user may improve the sales process, widening the potential market even for small firms or shops and lowering the TCO of show rooms by potentially minimizing the crowd and the waiting time for customers, and paves the way to more advanced technologies such as automatic shop assistants or sales advisors, personal shopper support systems or remote fashion advising.

In this paper we present the current state of development of vMannequin, a computer application designed to help fashion store customers to design the dress concept of their next purchase. The final goal of vMannequin is to be used in virtual fitting rooms, as an in-store solution, while the goal of the project is to obtain a flexible, computer based support for advanced virtual fashion applications. At the state, the focus is on enhancing the real time aspects of clothes dynamics, to increase the realism of the moving virtual mannequin.

The original contribution of this work is thus the description of the architecture of a customizable application that: i) allows the users to create, design and test new clothes on a virtual mannequin; ii) allows a data-driven configuration to customize it for different fashion contexts without requiring the writing of extra code; iii) provides a robust and innovative interaction model that can exploit available sensors and actuators to provide the final user with an immersive experience.

RELATED WORKS

Research and enterprise both have focused on virtual fitting rooms since more than a decade[1]. Many application that implements this idea are currently on market, sometimes very different one another in terms of goals and technologies involved. Some are conceived as plug-ins for e-commerce web sites, like e.g.: Virtusize [2], others as full web services like Fitnect[3]. In other cases still, like e.g. Fit.me[4], these web application leverages on robots able to simulate the size of the users. Augmented reality is enabled in case of products like Swivel [5], while triMirror [6] resorts on virtual avatars. However, it is difficult to find any technical details of these systems. For space constraints, we limit the related work analysis to those academic works closely related to ours. Dress dynamics in real time is considered by some authors, like e.g.: [7], where a physical based approach is used to realize real-time virtual try-on of garments by user interaction. Their approach is different from ours since the intended use of their application is to test dresses as they are designed by professional, moreover they do not specifically address animation of the virtual bodies in use. The introduction of the Microsoft Kinect sensor introduced novel interaction strategies, that are currently under investigation. In [8], the authors use an high definition camera to record the movement of the user, while the Kinect sensor analyze it. The analysis is then used to compute dynamic dress fitting which is then composed on the camera recoding. Even if the basic installation setting of the application is closely related to ours, their approach is not in real-time like ours. Moreover they do not address customization of dresses. The same overlapping technique is also exploited in [9], differently from our approach that preferred a virtual mannequin for the fitting. More recently, in [10] authors exploit the use of the virtual avatars for the fitting. This work lacks the visual appeal that is one of the focus of our proposed system, but introduced the use of real-time virtual body animation. Differently from it, however, we preferred an approach based on the recognition of a movement that triggers the closely matching animation present in a database of animation, instead of matching the user on time. In this way we avoided the animation artifacts that influenced their work.

THE SIMULATOR

The vMannequin simulator is intended to provide support to end-users involved in fashion concept design. The application will provide them with a variety of 3D assets, ranging from dresses to props, from shoes to hair styles that can be easily but thoroughly customized. A virtual 3D model, male or female as chosen by the user, can be dressed with the selected assets to show how they fit. The 3d model is displayed on a big screen, animated in real-time and can mimic users movements. To enhance the realism of the simulation, a great care has been devoted in implementing the visual part of the application using state-of-the-art shading techniques. Dress dynamics is also considered for the same reason.

In this section we describe the vMannequin simulator high-level architecture, highlighting the elements that compose the application and the environment where it is run, as well as their interaction. Figure 1 presents the application architecture, that the next Sections will describe into details.

Sensors

Essentially, two types of interaction are required to be handled. The first is the *customization interaction*, that is the interaction to customize the dresses before sending them to the 3D character for fitting. The second type concerns *animation interaction*, that is the sequence of gestures made by the user that causes the character to animate. Different currently on-market interaction devices can be adopted to these ends, ranging from QR-code reader or sound and motion recognition sensors to smart displays. We group them all under the definition of ‘sensors’ to focus on their ability to capture inputs from users and decode them into parametrized triggers for the application. The presence of different sensors is important since the customization interaction may require a deeper involvement as well as more precision gestures by the user, compared to animation interaction.

System Configuration

The simulator relies on an asset manager, for example a database, to handle the elements involved in visualization and customization of the the 3D models. This part of the application is intended to be transparent to the final user, but requires careful planning from the point of view of application developers. Since in this context developers requires not only programmers and IT technicians, but also 3D and computer graphics artists, as well as fashion and design experts which rarely have sufficient coding skills, the application has been developed to be *data-driven*. The goal is to have a simulator that can be completely configured, in term of types of assets available, sensors used and interaction models, by only inserting proper information in the database. This is to allow the maximum deployment flexibility in different kinds of fashion retail shops. After all, an high-end fashion store has different requirements in terms of customization options, animations, gestures to be recognized than a sport store.

The elements, or assets, required by vMannequin can be divided into four broad categories: Characters configurations, Animations, Gestures and Dresses. Since assets belonging to a category may need information stored for another category or needs to be related with it, a specific category of data is required to handle this inter-category communication: the Orchestration.

- *Characters configurations.* The virtual models to be used to show the dress fitting are essentially one male and one female 3D virtual model that can be adapted to the user’s needs. Features like weight, height, age, eye colors, skin tone, tattoos, nails colors can be customized as required and the database stores all the associated parameters. Depending on the degree of customization allowed, this category can require a larger or a smaller storage space.
- *Animations.* Virtual models are not static meshes but can be animated. Even if it is possible to let the 3D virtual models simply mimic the user’s movements this is not advisable. Animation directly mapped from a sensor’s body detection capability to the virtual model may result in visually awkward movements, that may break the simulation realism. Instead we propose an indirect mapping between the sensor readings and

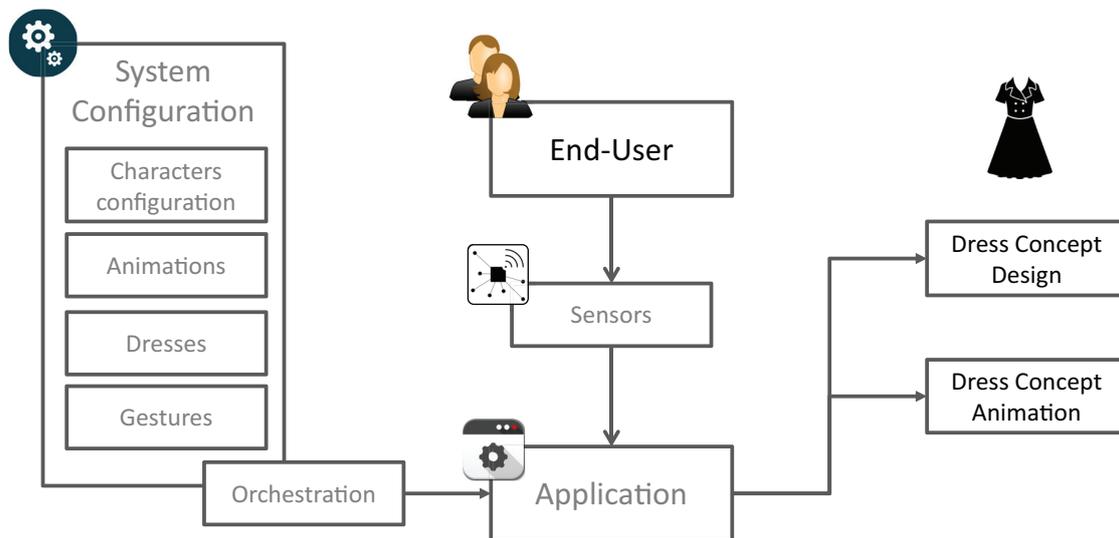


Fig. 1. The Simulator Structure

the animation played. The sensor recognizes an input pattern made by the user and the application select which of the database's available animations is the closest match. In this way, user's actions still affects the virtual mannequin in real-time but the realism of the movement is not compromised.

- *Dresses*. This category constitutes the bulk of the database, and again its size depends on the number of dresses, props, hairs the application is set to use. The meshes, i.e. the geometric definition of the 3D objects, can be further divided between *conforming* or *dynamic*. To the first group belong those dresses which follows skin-tightly the 3D virtual models (such as a pair of leggings). From a computational perspective they are the less expensive to handle because require the same animation techniques used for the virtual characters. Dynamic dresses, on the other hand, try to reproduce the majority of clothes physical characteristics and thus require specific management techniques, often resulting in an higher storage space requirement. In Figure 2 an example of the two kind of dresses is presented.
- *Gestures*. The interaction with the application is governed by gestures. Different type of sensors can allow different types of interactions. Full body sensors can recognize the position of the user, and return the orientations and the positions of the various joints that corresponds to the head, the torso, the arms and the legs. They can also recognize a sequence of movements as a specific action performed by the user. Speech recognition sensors might return identify words and sentences pronounced by the user and return. Pressure sensors or other haptic devices might return other interactions performed by the user. In this work, we will imagine that all the sensors will be able to identify a finite set of actions performed by the user. We will call *gestures* the data required to configure the sensors to identify the action performed by the user: for full body sensors, they correspond to sequence of



Fig. 2. The difference between conforming (a.) and dynamic (b.) clothes.

positions assumed by the user; for speech recognition sensors, they correspond to the vocabulary that must be recognized; and so on. Since the application might be configured to be used in different contexts (i.e. a sport goodies store, a bride-dress manufacturer, a department store), different gestures might be required

(a sport store customer might want to run or dance, while a bride-to-be might want to throw a bucket). The gesture data-base holds the specific gestures for the considered configuration. These gesture might trigger animation and configuration steps.

- *Orchestration*. The orchestration category holds the information required to connect the gestures, to the dress selection and configuration, to the characters and to the animation. Since the goal is to simplify the customization of the application using a data-drive approach, it exploits a formal specification (that will be described in next Section) that allows the setup of the interaction model without the requirement of programming skills.

THE APPLICATION

The purpose of the application is to enable the two objectives of the simulator: the design of a dress concept and the animation in real time of the result. To this end, it handles the following tasks: render the image on the big screen device, load and unload the assets from the database, allow the customization of these assets, react to the inputs received from interaction devices, play the animations. The rendering task requires a good trade-off between performances vs realism. One of the key factors of the vMannequin simulator is its ability to attract and involve shopper, hence the need for a state-of-the-art visual quality coupled with an immediate reaction to user's gestures. Visual quality can be easily achieved nowadays but comes with an higher requirement cost in terms of per asset storage space, which in turn translates in a possibly higher loading times before an element is displayed on screen. To this end the integration with the database is a critical issue, especially when considering the loading time of precomputed dynamic dresses. This is a non-trivial challenge that has been successfully addressed by producing a proprietary binary file format, which allowed, along with other techniques described in Section -A to reduce it considerably. In particular we have used principal component analysis based techniques to compress the animations, as well as a near-exhaustive precomputation of secondary cloth effects [11].

A. The database architecture

The database (which in this case can also be seen as a *file system*) contains all the assets required by the application as shown in Fig. 3. Note that the figure is not a classical entity-relationship diagram, since the configuration DB is not a relational database. In fact the configuration DB is a NoSQL database whose description goes beyond the scope of this paper.

To present the configuration DB we start with the table of the characters configuration component, that is represented with blue boxes. The *Character* box represents the type of characters that are available in the specific simulation. It usually includes two items (one for the male and another for the female), but can be increased depending on the context (for example to include young teens or children). Each element of the Character table also includes a small image that the application can use to show the preview of her selection. The *Geometry* box includes the meshes used by the application.

Each character must be connected to one and only one element of the geometry table. However the geometry table also holds data for other 3D objects that are used in the simulator such as dresses and add-ons which will be described later. Geometry data includes all the specifications required to properly draw the 3D objects: it includes a scene tree composed of several hierarchically interconnected nodes and mesh pointers, a set of index buffers, a set of vertex buffers. For characters and conforming clothes it also includes a bone hierarchy and a binding pose expressed using offset matrices [12]. Vertices have a variable format that always includes the positions, and accompany it with the directions of vertex normals, the UV coordinates. For characters and conforming clothes vertices includes also a set of up to four indices to influencing bones, and the weight using to blend final pose. The character configuration also includes the *Texture* box that, as the name suggests, includes all the texture sets associated to a character. Texture elements are collections of several images used by the shader to produce the final render: they usually includes diffuse color map, transparency map, normal and bump map and specular reflection map. Each character can be associated to more than one texture to allow simple customizations like changing the tone of the skin or the color of the eyes. Each texture set is thus also characterized by a preview image. In special circumstances characters might not have any texture involved: in this case the application will allow the user to select a fixed color.

The dresses components are represented in Fig. 3 with green boxes. The *Dress* table contains one element for each dress supported in the configuration. Again, each element of the dress table also includes a small image to preview its appearance. As for characters, each dress must have a pointer to an element of the geometry table and might have a pointer to one or more texture sets. As previously introduced, dresses are divided into two main types: conforming and dynamics. However, the dress table includes also a third type of element, called *body features*. The latter is used to add body features like hairs, piercing or tattoos. Body features can either be conforming or dynamic (to support dynamic hairs). In case of tattoos, the features might not have associated a geometry. Body features are implemented by the simulator as normal dresses: however the distinction allows the introduction of the elements in different locations of the user interface, and prevent them to be considered as clothe features to be exported to the real dress production plan. Not all dresses could be tailored for all characters (e.g. a kids clothe cannot be fit on an adult): for this reason the table *Can fit* represents a many-to-many relation that defines which clothes can be worn by a given character. Dresses are also grouped into different types, as defined by the table *Clothe Type*. In this way the application can present the clothes grouped into different categories, considering for example tops, trousers, leggings, socks, shoes and so on. The table *Clothe clash* contains instead couples of clothe types that cannot be worn together. It avoids for example to fit at the same time two evening dresses, one on top of the other. Beside changing the texture, clothes can also be configured as specified in the table *Add-on*. Each dress might have associated zero or more add-ons, which are divided into two types: *Decals* and *Props*. As for other customizable elements, they are both characterized by a preview image. Decals are simple texture overlays that can be superposed to the fabric to

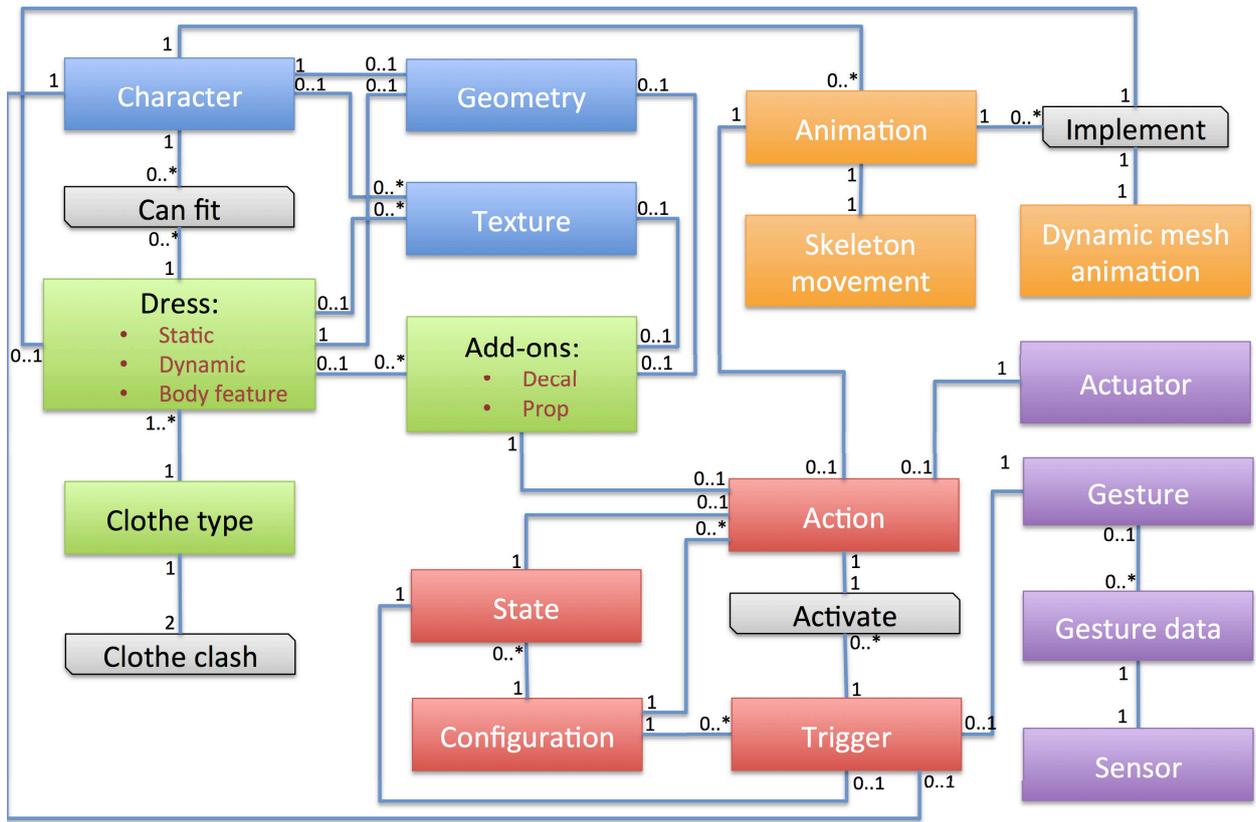


Fig. 3. The Application Database: character configurations (blue), animations (orange), dresses (green), gestures (purple) and orchestration (red).

create an high level, but controlled, degree of personalizations. The application allows to move, rotate and scale the decal on the surface of the dress. Since decals are basically extra textures, they are connected to one and only one element of the Texture table. Props are instead pieces of geometry that can be superposed to the dress (such for example an extra button or a pendant). The application allows to position and rotate them, keeping them anchored to the surface of the dress. Props requires a pointer to one and only one element of the geometry table, and can point to one or more elements of the texture table to allow the selection of different texture sets to further customize the add-on.

Elements of the animation component of the DB are represented with orange boxes in Fig. 3. In particular, each animation that a character can perform has a corresponding element in the *Animation* table. Each animation is characterized by its length in frames, and it must be associated to a skeletal animation that is stored into table *Skeleton movement*, which holds the positions, rotations and scaling of all the bones associated with a character. Dynamic clothes must also have associated an element of the *Dynamic mesh animation table* that stores the compressed deformations of the dress to produce a realistic effect. Elements of table *Implements*, associate the deformation to a given animation for a specified dynamic dress.

The gesture component of the DB is shown with purple boxes in Fig. 3. Table *Sensor* and table *Actuator* include respectively the description of all the sensors and actuators connected to the application. Each of the element of both tables

select among a possible set of sensors (i.e. a Microsoft Kinect, a Midi device, a step-up motor, a LED based colored dynamic illumination system) preconfigured in the application, that can be used in the specific configuration. Elements of the *Gesture* table hold instead the possible user interactions that can be identified by the application: they can for example correspond to the detection of the wave of the arm, or the movement of a joystick, the interaction with central-left portion of a smart screen and so on. Each gesture must be then defined through a set of sensor-specific parameters contained in the *Gesture data* table. Note that the application defines a special type of sensor/actuator: the *timer*. Timers are special actuators that can be started, and produces a sensor reading when a given time elapses. This can be used to trigger special actions for example when the user does not interact with the application for a prolonged time.

In Fig. 3, red boxes represents the orchestration part of the DB. In particular, the application can run in one of several configurations, each one represented by an element of the *Configuration* table: this may allow the shop owners to run the installation in different operating modes, depending on the expected affluence in number and type of customers. Each configuration is characterized by a finite set of states stored in the *State* table. This allows to use state-machines as a theoretically proven effective mechanism to store the cause-effect interactions of the configuration in a data-driven way. In particular, the application can perform a set of actions that are defined in the *Action* table. An action can either activate

an actuator, play an animation or change the current state (as shown with the possible connections with the elements of the corresponding tables). Actions are started by an element of the *Trigger* table. Triggers are fired whenever the associated gesture is detected. The effect of the trigger can be confined to be effective only in a given state, or for a given character. Finally, each trigger can fire more than one action, as specified by the elements of the *Activate* table.

Users and their Experience

In Figure 4 it is possible to preview the expected installation of vMannequin in a fashion store. The user is detected by the application (In Figure 4-A) which exits its idle status and becomes active. A short set of instructions can be provided to the user at this point. If the user decides to interact with the simulator, it is prompted by the request for the specification of some characteristics, like gender, height, size, age, for the virtual 3D model used for the fitting. This step can be substituted by automatically detecting the user's features but this may introduce more problems than not. What if the dress to be purchased and that the user wants to virtually fit is not meant for her but as a present for someone else?

After this first stage, a catalog of the available customizable assets (mainly dresses, but shoes, hair styles are available too) is presented to the user (In Figure 4-B). The user is supposed to interact with the application for as long as required to be satisfied with the options selected. Each time an asset is completed, it can be sent to the 3D character and seen fitted. The user is not required to completely dress the character, and can see the real-time fitting and animation at any moment.

In the last stage, the 3D character is dressed up to the point the user needs (In Figure 4-c). At this point her experience is concluded and the results of its interaction, both in terms of pictures, both in terms of a checklist with the chosen dresses and their customization can be sent to her mobile, mail account or other preferred communication methods.

PROOF OF CONCEPT

In this Section we present the proof-of-concept application that implements the most relevant features introduced in the previous Section. In Figure 5 the architecture of this system is shown. The goal of the application is to verify the technical feasibility of the proposed model.

The application has been developed using Microsoft Visual C# [13], while users can interact with the simulator by means of a Microsoft Kinect II for XBOX ONE sensor [14]. The sensor offers a wide range of detection features that can be used to implement both the customization and the animation interaction types. Other interaction devices are currently under study, since the customization procedure, because of its specific needs, can be demanded to a different device such as a smart screen or a mobile device connected to the application. The same big screen on which the simulation result is displayed can be also 'smart' to allow the customization of interactions.

Rendering is performed using a specifically developed rendering engine tailored on the specific requirements of the simulator, in terms of performances vs realism, and uses simple

but effective three points light model to produce believable images. To allow the maximum deployment flexibility on different hardware, the OpenGL graphic library has been chosen, accessed through the OpenTK [15] wrapper. Since it is widely supported by different display adapters, the use of OpenGL could be of benefit especially in case of a future integration with mobile displays.

In order to have high quality models for the asset manager, we leveraged the repository of Daz 3D, a software dedicated to morphing, posing, animating virtual characters. The general idea is to use easy-to-find detailed objects that can near as much as possible the clothes that the retail store wants to show with the application. With a little extra effort by modeling professionals, those objects can be customized to exactly match the clothes in stock. Currently there are different similar programs, e.g. Poser, Bryce, MakeHuman, so there is a high number of available dress models, hair styles, props and other elements that can be reused.

Because of this, the origin of the used 3D models can be very different, introducing the challenge of having many different file formats to handle for the asset manager. This has been solved by using the Assimp[12] library, able to import in a uniform manner a vast number of standard formats. Since Assimp is currently unable to read dynamic dress information, stored as mesh *morphs targets* (or *blend shapes*) that only two file formats are able to save (i.e.: Collada *.dae and autodesk *.fbx), we used Autodesk FBX SDK 2015.1 [16] to import them. 3D Model files decoded using these library are transformed in the proprietary format used by our application. Models are transformed off-line, before being used by vMannequin simulator. Since performances are an important issue for the concern of assets management (the user should experience immediate visualization of the chosen clothes and the applied customizations), the development of a proprietary binary file format allowed us to significantly shorten the loading time of dresses, especially of the precomputed dynamic ones. The animations used in the example application were baked at 30fps, and the simulator can run them while displaying one animating character in full attire, with at least one dynamic dress fit on it. The test was run on a machine equipped with an Intel Core i7 2.4GHz, 8GB of RAM and a NVIDIA GeForce GT670M at a 1920 × 1080 resolution.

CONCLUSIONS

In this paper we presented the current state of development of vMannequin, a computer application designed to help fashion store customers to design the dress concept of their next purchase. Future works goes in the direction of improving the performances of the application, also to adapt it to low end architectures. Moreover, we aim to enhance the realism of the simulated environment by means of techniques like spherical harmonics, that allows the illumination of the 3D models to match that of the retail store, to achieve an higher degree of realism and immersiveness. Next development stage will be directly tested by users to improve the interaction part, as well as its ability to improve retail store attractiveness.

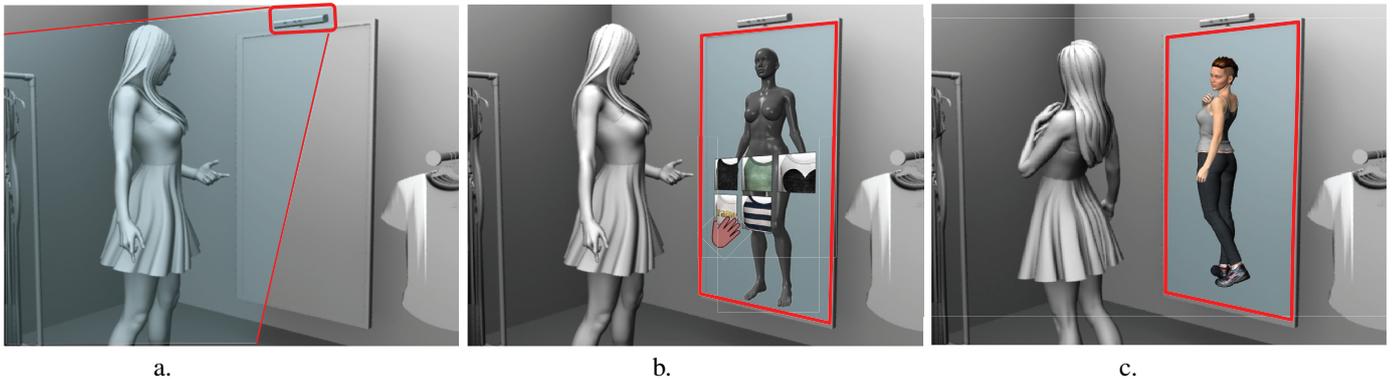


Fig. 4. The user experience. The Kinect sensor detects a user standing in front of the installation (a). The user interacts with the application, by Kinect or smart screen (b). The user can simulate the fitting of the dress concept created (c).

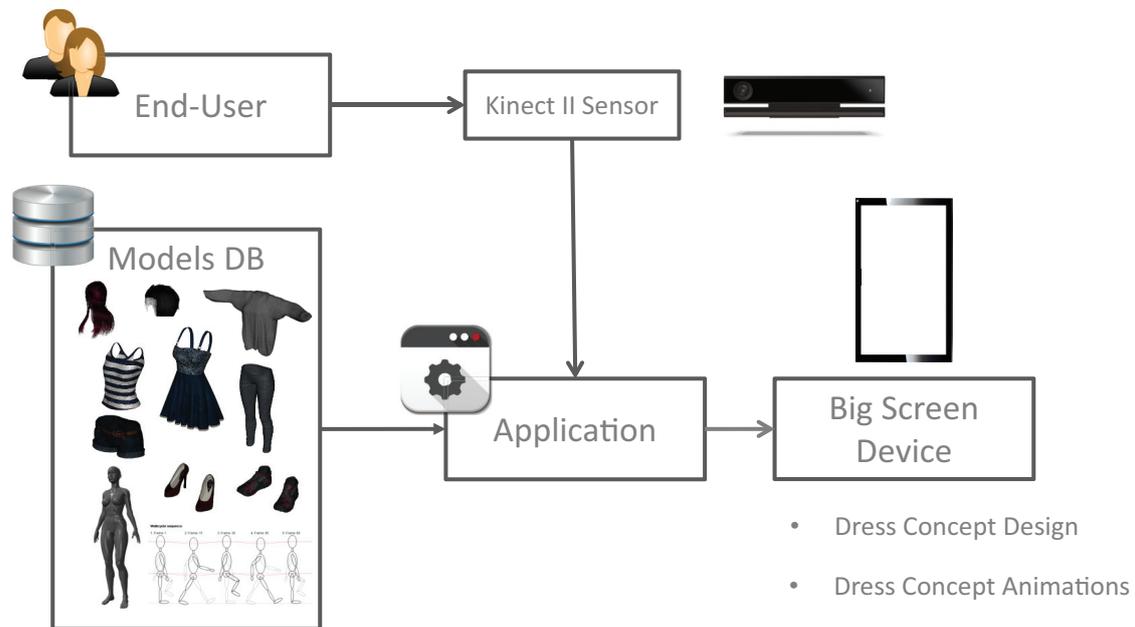


Fig. 5. The Simulator Proof-of-Concept.

REFERENCES

- [1] D. Protopsaltou, C. Luible, M. Arevalo, and N. Magnenat-Thalmann, *Advances in Modelling, Animation and Rendering*. London: Springer London, 2002, ch. A body and Garment Creation Method for an Internet Based Virtual Fitting Room., pp. 105–122. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-0103-1_7
- [2] Virtusize fitting solution. [Online]. Available: <http://www.virtusize.com/site/>
- [3] Fitnect 3d fitting room system. [Online]. Available: <http://www.fitnect.hu/>
- [4] Fit.me website. [Online]. Available: <http://fits.me/>
- [5] Swivel virtual try-on system. [Online]. Available: <http://www.facecake.com/swivel/>
- [6] trimirror virtual fitting room. [Online]. Available: <http://www.trimirror.com/en/about/>
- [7] Y. Meng, P. Y. Mok, and X. Jin, “Interactive virtual try-on clothing design systems,” *Comput. Aided Des.*, vol. 42, no. 4, pp. 310–321, Apr. 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.cad.2009.12.004>
- [8] S. Giovanni, Y. C. Choi, J. Huang, E. T. Khoo, and K. Yin, *Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Virtual Try-On Using Kinect and HD Camera, pp. 55–65. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34710-8_6
- [9] S. Hauswiesner, M. Straka, and G. Reitmayr, “Virtual try-on through image-based rendering,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1552–1565, 2013.
- [10] U. Gltepe and U. Gdgbay, “Real-time virtual fitting with body measurement and motion smoothing,” *Computers & Graphics*, vol. 43, pp. 31 – 43, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0097849314000600>
- [11] D. Kim, W. Koh, R. Narain, K. Fatahalian, A. Treuille, and J. F. O’Brien, “Near-exhaustive precomputation of secondary cloth effects,” *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 87:1–7, July 2013.
- [12] A. Gessler, T. Schulze, and K. Kulling. The open asset import library. [Online]. Available: http://assimp.sourceforge.net/main_doc.html
- [13] C# reference for visual studio 2015. [Online]. Available: <https://msdn.microsoft.com/en-us/library/618ayhy6.aspx>
- [14] Kinect for windows software development kit (sdk) 2.0. [Online]. Available: <https://dev.windows.com/en-us/kinect>
- [15] OpenTK toolkit library. [Online]. Available: <http://www.opentk.com/>
- [16] Fbx data exchange technology. [Online]. Available: <http://www.autodesk.com/products/fbx/overview>

A SOFTWARE FRAMEWORK FOR INTELLIGENT COMPUTER-AUTOMATED PRODUCT DESIGN

Robin T. Bye*, Ottar L. Osen*,† Birger Skogeng Pedersen*,
Ibrahim A. Hameed*, and Hans Georg Schaathun*

* Software and Intelligent Control Engineering Laboratory
Faculty of Engineering and Natural Sciences
Norwegian University of Science and Technology
NTNU in Ålesund, Postboks 1517, NO-6025 Ålesund, Norway

† ICD Software AS

Hundsværgata 8, NO-6008 Ålesund, Norway

KEYWORDS

Virtual Prototyping; Product Optimisation; Artificial Intelligence; Genetic Algorithm

ABSTRACT

For many years, NTNU in Ålesund (formerly Aalesund University College) has maintained a close relationship with the maritime industrial cluster, centred in the surrounding geographical region, thus acting as a hub for both education, research, and innovation. Of many common relevant research topics, virtual prototyping is currently one of the most important. In this paper, we describe our first complete version of a generic and modular software framework for intelligent computer-automated product design. We present our framework in the context of design of offshore cranes, with easy extensions to other products, be it maritime or not. Funded by the Research Council of Norway and its Programme for Regional R&D and Innovation (VRI), the work we present has been part of two separate but related research projects (grant nos. 241238 and 249171) in close cooperation with two local maritime industrial partners.

We have implemented several software modules that together constitute the framework, of which the most important are a server-side crane prototyping tool (CPT), a client-side web graphical user interface (GUI), and a client-side artificial intelligence for product optimisation (AIPO) module that uses a genetic algorithm (GA) library for optimising design parameters to achieve a crane design with desired performance. Communication between clients and server is achieved by means of the HTTP and WebSocket protocols and JSON as the data format.

To demonstrate the feasibility of the fully functioning complete system, we present a case study where our computer-automated design was able to improve the existing design of a real and delivered 50-tonnes, 2.9 million EUR knuckleboom crane with respect to some chosen desired design criteria.

Our framework being generic and modular, both client-side and server-side modules can easily be extended or

replaced. We demonstrate the feasibility of this concept in an accompanying paper submitted concurrently, in which we create a simple product optimisation client in Matlab that uses readily available toolboxes to connect to the CPT and optimise various crane designs by means of a GA. In addition, our research team is currently developing a winch prototyping tool to which our existing AIPO module can connect and optimise winch designs with only small configuration changes. This work will be published in the near future.

INTRODUCTION

With the geographical heart of the Norwegian maritime cluster located on campus, NTNU in Ålesund (formerly Aalesund University College) has played the role of a hub for collaborations in education, research, and innovation. According to the Global Centre of Expertise (GCE) Blue Maritime, one of only three industrial clusters in Norway that have been awarded this prestigious title and funding from the Norwegian Innovation Clusters Programme, the Norwegian maritime cluster consists of 13 design companies, 14 ship yards, 20 ship-owner companies, 169 equipment suppliers, 22,500 employees, and a total annual revenue of about 5.7 billion EUR.¹ Together with two of these companies, ICD Software AS (provider of industrial control systems software)² and Seaonics AS (designer and manufacturer of offshore equipment),³ we have received funding from the Research Council of Norway and its Programme for Regional R&D and Innovation (VRI) for two independent but related research projects (grant nos. 241238 and 249171) for using artificial intelligence (AI) for intelligent computer-automated design (CautoD) of offshore cranes and winches, respectively.

Our main focus is on the development of a generic and modular software framework for intelligent CautoD of products such as crane and winches. We use the recently completed crane design project for exemplification but emphasise that our work has strong synergies with our

Corresponding author: Robin T. Bye, robin.t.bye@ntnu.no.

¹www.bluemaritimecluster.no, accessed 8 February 2016.

²www.icdsoftware.no

³www.seaonics.com

ongoing winch design project and generically can be extended to other products and CautoD methodologies.

In the following, we begin with a background overview of virtual prototyping (VP) in general and CautoD in particular, offshore cranes, and the motivation and aim of our work. Next, we outline the method we have used, including details about the software architecture of our framework and each of its main components. Then we demonstrate how our first, complete version of our system is able to improve the design of a real knuckleboom crane that has already been sold and delivered to a company in Baku, Azerbaijan. Finally, we discuss our results and impact on future work.

BACKGROUND

Virtual Prototyping (VP)

We may define VP as the computer-aided construction of digital product models (usually virtual prototypes or digital mockups) and realistic graphical simulations for the purpose of design and functionality analyses in the early stages of the product development process (Pratt, 1995). Both the shipbuilding (Kim et al., 2002) and automotive (Wöhlke and Schiller, 2005) industries have made significant use of VP aspects such as modelling, simulation, and visualisation techniques, in the process of evaluating and improving product design and to the validation of product planning and manufacturing processes (e.g., Mujber et al., 2004; De Sa and Zachmann, 1999; Weyrich and Drews, 1999).

Common VP methodologies include computer-aided design (CAD), realistic virtual environments (VEs), virtual reality (VR), and CautoD. CAD is related to the use of computer systems to aid in the creation, modification, analysis, or optimization of a design (Narayan et al., 2008), and is generally associated with 2D and 3D modelling software. VEs can be used to improve collaboration and teamwork in product design, for example in construction engineering (e.g., Waly and Thabet, 2003; Sarshar et al., 2004; Yerrapathruni, 2003) or component and assembly design (e.g., Shyamsundar and Gadh, 2002; Chang et al., 1999; Chan et al., 1999), and can be particularly useful in multidisciplinary product development (e.g., mechatronics engineering), to avoid inefficient communication between designers and engineers with different backgrounds that can slow the design process (Shen et al., 2005). Likewise, employing VR as a collaborative VP tool can greatly improve the product design, test and review loop before committing to physical fabrication (Choi and Cheung, 2008). Here, our main focus is on applying AI for CautoD, and a GA in particular, to automate and optimise the design phase of product development.

Computer-Automated Design (CautoD)

Likely the first scientific report of CautoD was reported by Kamensky and Liu (1963), who created a computer programme for determining suitable logic circuits satisfying certain hardware constraints while at the same time evaluating the ability of the logics to perform character

recognition. Many contributions of CautoD have been made since then, particularly in the field of structural engineering (see Hare et al., 2013, for a survey). The general paradigm is that of optimisation, where one formulates the design problem as the optimisation of an objective function; that is, minimisation of a cost function or maximisation of a fitness function. For complex optimisation problems for which analytical and exact solutions are difficult or impossible to obtain, AI methods such as machine learning and evolutionary computation can often be used with great success (see Zhang et al., 2011, for a survey).

A seminal example of design optimisation using AI is the simulated annealing (SA) algorithm (Kirkpatrick, 1984; Černý, 1985), which is still in use today, e.g., for CautoD of tensegrity systems (Xu and Luo, 2010). Another highly influential algorithm is the GA, which has been used for a variety of design optimisation problems, including computer-based control of gas pipeline systems (Goldberg, 1983), structural systems (Rajeev and Krishnamoorthy, 1992), 2D geometrical nonlinear steel-framed structures (Pezeshk et al., 2000), the piping process of offshore drilling platforms (Peng et al., 2010), and the influential work on design planar and space structures by Erbatur et al. (2000). Finally, we would like to guide the reader to the work of Kaveh and Talatahari (2009), who used particle swarm optimisation (PSO) and ant colony optimization (ACO) to find an optimal design of different types of truss structures, and the work of Schoning and Li (2013); Schöning (2014), who used a set of different AI algorithms that could connect and interrogate a simulator for a homogeneous charge microwave ignition system in search of novel design solutions, much in the same vain as the software framework we present here.

Virtual Prototyping of Offshore Cranes

An offshore crane such as the one in Figure 1 is a complex system of components interacting to achieve safe movement of heavy goods, often under harsh and difficult conditions.

Even simple versions of such offshore cranes consist of a large number of components, including hooks, winches, slewing rings, cylinders, booms, hinges, sheaves, and pedestals (see Figure 2 for an illustration of the main components). The choice of crane components and their physical properties and interrelationships determines various measures of performance of interest to the crane designer. For example, key performance indicators (KPIs) such as the desired workspace, the working load limit (WLL) and the safe working load (SWL) within that workspace; the total weight of the crane; its control system characteristics, durability, installation and operating costs; and safety concerns are all factors that the crane designer must take into account when designing a new crane. In addition, laws, regulations, and the use of design codes such as the standards provided by classification societies like DNV-GL, Lloyd's Register Group Limited, and the

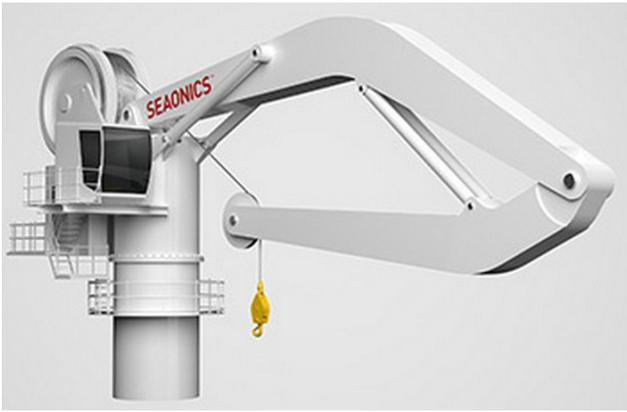


Figure 1: Boomerang-shaped knuckleboom crane by Seanonics AS with an ingeniously designed wire or fibre rope running directly from winch to boom tip. The design increases the workspace compared to standard knuckleboom cranes while reducing wear and tear of the wire or fibre rope. This crane is ideal for arctic operations and fibre rope use. Image courtesy of Seanonics AS.

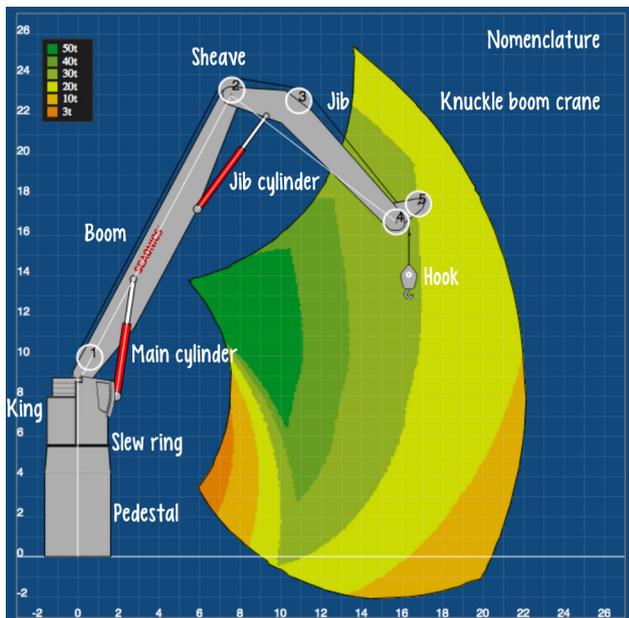


Figure 2: Illustration of the main components of an offshore knuckleboom crane and its 2D load chart. Image courtesy of ICD Software AS.

American Bureau of Shipping all put constraints on the choice of design parameters.

VP is currently the main focus of research of the Software and Intelligent Control Engineering (SoftICE) Laboratory⁴ at NTNU in Ålesund. For the last few years, the SoftICE lab and our colleagues in the Mechatronics Laboratory⁵ have published several papers relating to VP and aspects of modelling, simulation and visualisation of offshore cranes, marine operations, installations, and industrial robotics (e.g., Bye et al., 2015; Sanfilippo et al.,

⁴blog.hials.no/softice

⁵www.mechatronics.hials.org

2013, 2015; Chu et al., 2014, 2015; Chaves et al., 2015; Hatledal et al., 2015).

Of particular interest is the recent work by colleagues Hatledal et al. (2015), who present a voxel-based numerical method for computing and visualising the 3D workspace of offshore cranes. Despite the importance of a crane's lifting capacity in different positions in the workspace (often visualised as a 2D load chart), which depends on the properties of crane components such as the length of the boom and jib and the maximum pressure of their cylinders, workspace and load chart calculations are usually not taken into account in the design phase and are merely realised as an indirect consequence of a priori design choices (Hatledal et al., 2015). However, Hatledal et al. note that employing their algorithm as a trial-and-error VP tool during the preliminary design phase, factors such as the length of the boom and jib and the size of their cylinders can be designed to yield better workspace characteristics.

Several other researchers have recently published relevant work relating to VP of offshore cranes, operations, and control. Bak et al. (2011); Bak and Hansen (2013); Peng et al. (2010); Pawlus et al. (2014) present VP of offshore knuckleboom cranes and pipe handling equipment, including techniques for modelling, simulation, and parameter identification that can aid in the VP process. Park and Le (2012) employ VP technology for modelling and control design of a mobile harbour crane system, whereas He et al. (2014) focus on VP-based multibody systems dynamics analysis of offshore cranes.

The work presented above shows that VP of offshore cranes and is an active field of research, however, as we will detail in the following section, most of this work can be improved by intelligent automation of the design process, which is the motivation for our own work.

Motivation and Aim

While the contributions above are valuable, the various models, calculations, simulations, and visualisations are mainly used as VP tools requiring a human to manually try and test design solutions until satisfied. For example, the work of Hatledal et al. (2015) describe how their VP environment can be used for a trial-and-error procedure to improve the traditional experience-based rule-of-thumb approaches employing pen-and-pencil or spreadsheet calculations commonly employed in the maritime industry but their method is hardly satisfactory given the large number of design parameters involved.

The findings reported in the literature typically describe various means to determine offshore crane properties and behaviour based on pre-determined design parameters, analogous to calculating the forward kinematics of a robotic arm. However, the inverse problem, namely that of choosing appropriate, possibly conflicting, values for numerous offshore crane design parameters such that some desired design criteria are satisfied, is much harder.

The aim of our two research projects is to solve this problem for respectively offshore cranes and winches

by means of an intelligent CautoD software framework employing a GA for searching through the vast number of design choices and combinations until an optimised design is found. The design will inevitably involve tradeoffs and thus be Pareto optimal, which means that improving some aspect of the design will necessarily detract from another.

METHOD

Software Architecture

The diagram in Figure 3 shows the software architecture for our complete system.

We have adopted a server-client software architecture, in which the CPT act as as server to which clients such as the AIPO module and the web GUI can connect via two different communication interfaces. The first interface uses the Hypertext Transfer Protocol (HTTP), with data transferred as JavaScript Object Notation (JSON), a lightweight human-readable data-interchange format. The second interface uses WebSocket (WS), which is a protocol providing full-duplex communication channels over a single TCP connection, also with JSON as the data format. Because WS enables streams of messages on top of TCP, using WS for communication is advantageous for bidirectional conversations involving many small messages being sent to and from a server. If this is not a concern, clients may prefer to use HTTP, since this protocol is very mature and well supported in many different programming languages. Two server modules that implement the communication interfaces both connect to a third module, the crane calculator. All three modules are implemented in Java and together constitute the CPT.

On the client side, we have implemented an AIPO module that uses a GA library module for optimisation. Both modules are programmed in Haskell and are used for CautoD of offshore cranes. In addition, we have implemented a web GUI in JavaScript. Finally, to test the modularity of the software framework, we have also implemented a crane optimisation client in Matlab. The Matlab client and a test suite for optimal crane design is presented in an accompanying paper submitted together with this paper (Hameed et al., 2016).

The CPT server accepts messages that conform to a set of rules that we have defined. Specifically, the AIPO module sends a JSON message object consisting of three parts (subobjects): (i) a "base" object with a complete set of default design parameter values; (ii) a "mods" object with a subset of design parameter values that modifies the corresponding default values; and (iii) a "kpis" object with the desired KPIs to be calculated and returned by the CPT.

The design is highly modular. On the server side, we can swap prototyping tools as long as they conform to the HTTP/JSON or WS/JSON communication interfaces and message formats that we have defined. Thus, we can easily create a winch prototyping tool (currently under development in a parallel project) or another product prototyping tool, as long as the tool is able to receive such messages as described above and calculate and return desired KPIs. Likewise, on the client side, different optimisation clients

can connect to the CPT (or another product optimisation tool), again as long as they conform to the communication interface.

Crane Calculator

The components of an offshore crane may total several thousand parameters, making it infeasible to manually pick good values for each parameter. However, through the years, crane designers have been able to reduce this number to a set of about 120 design parameters that are considered the most important. Based on the values of these parameters, which can be set manually or by a CautoD tool such as AIPO, our crane calculator is able to calculate a fully specified crane design and its associated KPIs. The goal of the designer is thus to determine appropriate design parameter values that achieve desired design criteria (KPIs), while simultaneously meeting requirements by laws, regulations, codes and standards.

The accuracy of our crane calculator has been verified against other crane calculators and spreadsheets currently in use in the industry, and, as a result, Seonics AS has already adopted the CPT server and web GUI client for manual crane design.

A block diagram depicting how the crane calculator can be used for manual crane design is shown in Figure 4.

Web Graphical User Interface (GUI)

To simplify practical use of the crane calculator, we have created a web graphical user interface (GUI) that can be used to interact with the crane calculator via WS/JSON communication. Using the web GUI to manually adjust the 120 design parameters in the crane calculator by trial-and-error, the effect of the parameters on a number of KPIs and other design criteria can be investigated numerically, with the possibility for exporting to text files, and visually, by depicting the main components of the crane and its 2D SWL load chart.

The GUI consists of three parts aligned horizontally: a left column with "drawers" where design parameters can be modified; the visualised crane and load chart in the middle; and a right column that shows numerical results, including the load vector (position and SWL), slewing ring torque, boom angle, jib angle, and the main and jib cylinder data (compression force, buckling force and SWL, and pressure and its SWL).

The load chart is a graphical representation of the lifting capacity in the workspace of a particular crane design as determined by the crane calculator. The workspace is divided into zones, where each zone is highlighted by a colour indicating the maximum SWL in tonnes in that particular crane configuration.

Due to space consideration, we refer to Bye et al. (2015) for a screenshot of the GUI.

Artificial Intelligence for Product Optimisation (AIPO)

The manual design process using the web GUI together with the CPT is cumbersome. Indeed, there are more

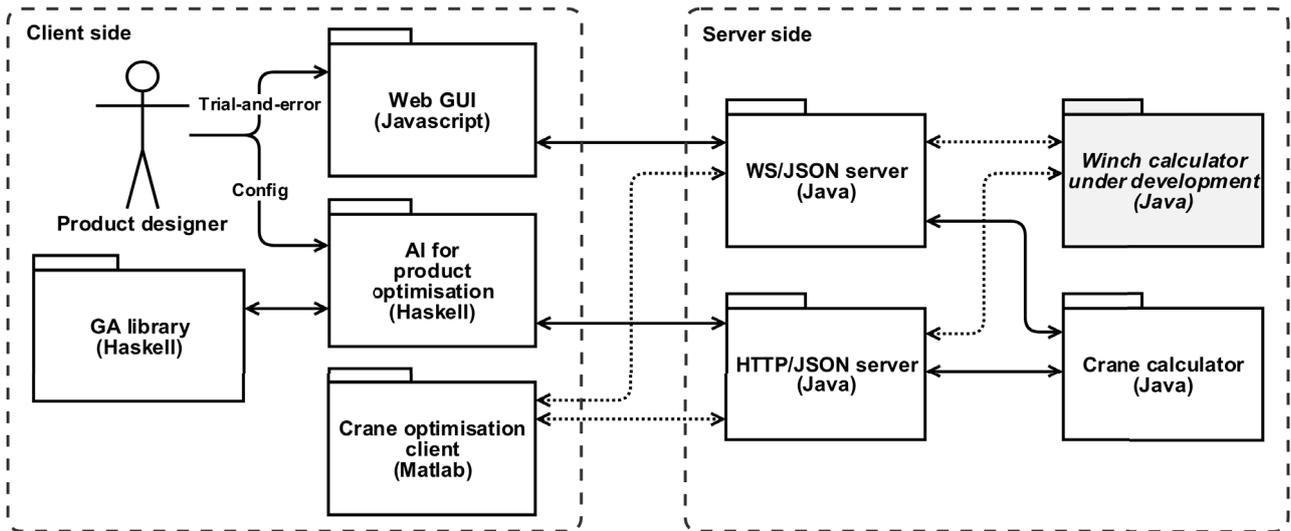


Figure 3: Software architecture for intelligent CautoD of offshore cranes, winches, or other products. The winch calculator is shown in grey because it is still under development. Solid and dashed lines indicate whether modules and their interconnections are inside or outside the scope of this paper, respectively.

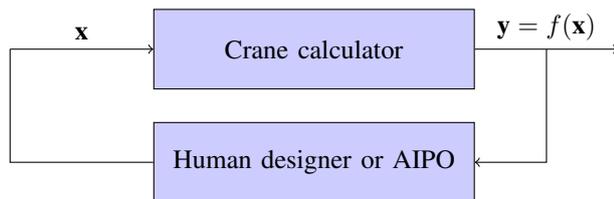


Figure 4: Human crane designer using a manual trial-and-error approach with the crane calculator to tune the input design parameters \mathbf{x} until the resulting design \mathbf{y} matches the desired design criteria. Our aim is automate this process by a CautoD solution, namely AIPO.

than 120 parameters that must be specified by the crane designer. Clearly, this large number of parameters makes the search space (the space of all possible combinations of parameter values) very large and a manual trial-and-error approach will necessarily be both time-consuming and cost-inefficient and lead to suboptimal designs. Hence, we have developed an AI for product optimisation (AIPO) software module replacing the human operator in Figure 4 in order to automate and optimise the design process.

The product designer must configure the AIPO module with one or several objective functions based on the KPIs and design criteria of the product to be designed, be it an offshore crane, a winch, or some other product. In addition, the product designer must configure the AI algorithm to be used for optimisation. For example, if using a GA, the designer must set certain parameter values such as population size, mutation rate, and choose methods for selection and crossover.

Using AI optimisation methods such as a GA in this case, the module interrogates the CPT with an objective function until an optimised design solution is obtained. A short description of GAs and objective functions are

provided in the following sections.

Genetic Algorithms (GAs)

The GA is inspired by natural evolution, with elements such as inheritance, mutation, selection, and crossover. It is well suited for hard optimisation problems where solutions are difficult or impossible to obtain in polynomial time. Additionally, an advantage of GAs is that constraints can be handled with ease. For crane design, the most important components are the hydraulic cylinders and the slewing ring. In the case of the slewing ring, the maximum torque is of special interest, because it limits both the lifting distance and weight of the payload. In the case of the cylinders, their maximum pressure and buckling limit is of major importance. These and other constraints can easily be incorporated in a GA optimisation procedure.

Most of the literature attributes the GA to Holland (1975), with subsequent popularisation by Goldberg (1989), and it is currently a very popular optimisation tool across many different disciplines (e.g., Haupt and Haupt, 2004).

The authors and colleagues have used GAs for a number real-world optimisation problems (e.g., Bye, 2012; Bye and Schaathun, 2014, 2015; Bye et al., 2015; Alaliyat et al., 2014; Sanfilippo et al., 2013). Here, we let the AIPO module use a GA in conjunction with the CPT for optimising the design phase of offshore cranes. Nevertheless, with only minor adjustments, we are able to adopt the same process for CautoD of winches (currently under development) or other products. More details about the GA we use here can be found in a previous paper (Bye et al., 2015).

Objective Functions

In order for the GA to succeed in optimising design solutions, a suitable objective function must be selected

that incorporates the design criteria that we wish to optimise. That is, the GA must determine values for a number of input parameters \mathbf{x} such that the selected objective function $f(\mathbf{x})$ is optimised.

Nevertheless, choosing an appropriate objective function is not trivial. In addition to adhering to laws, regulations and standards, offshore cranes must be designed in accordance with the specific needs of the customer. Optimising such a set of potentially conflicting design criteria (that is, there is a tradeoff between two or more objectives) is called multiobjective optimisation (MOO). Using a GA for MOO, the GA will not return a single solution but a set of Pareto optimal solutions, which means that none of the objective functions can be improved without degrading others (e.g., see Haupt and Haupt, 2004; Arora, 2012, for details).

Choice of Implementation Languages

The web GUI was implemented in JavaScript, the CPT in Java, and the AIPO module and the GA library in Haskell.

JavaScript has stayed popular for more than a decade and has a huge user base and easy access to resources. Most web client supports JavaScript out of the box. Since only a text editor and web browser is needed for creating programs, modifications can be tested immediately by refreshing the browser, thus enabling rapid development and testing. Initially, therefore, the crane calculator was part of a JavaScript web client, but this choice had several drawbacks. Most importantly, the source code was open for everyone to see, which was unacceptable to our industrial partners. In addition, processing speed was dependent of the client's hardware. Thus, to ensure sufficient processing power independent of user hardware, the JavaScript implementation was moved "as is" to the server-side. Unfortunately, there were speed issues and further development was needed in order to develop a suitable application programme interface (API) for the AIPO module. We therefore decided to opt for a different implementation language able to solve these issues, and also chose to separate the calculator from the handling of communication with clients. In order to keep migration costs low, a language with syntax similar to the C-like syntax of JavaScript was desired. Java was therefore chosen, with the additional advantage of its portability to different platforms. If even higher performance will be needed in the future, we might use compiled languages such as C, C++ or C#, or a functional language like Haskell, which is well suited for parallelisation.

For the AIPO module and its interconnected GA library, we chose Haskell. Haskell is a purely-functional programming language, which means that functions in Haskell are pure, there is no global state, and no side effects. In addition, the separation between pure and impure functionality makes code easier to debug. Code written in Haskell is therefore less error-prone and usually more concise, compact, and readable than imperative programming languages like C or Java.

Haskell is a good choice for parallel programming, which we believe is likely to be needed as the complexity of our software framework grows. Using pure parallelism guarantees deterministic processes and zero race conditions or deadlocks, however, non-pure concurrency related to pseudorandom number generators and other processes is still required.

Using Haskell for AIPO and GA implementation makes this part of our framework very modular and extensible, something we believe is necessary in order to expand the framework and design tools in the future.

Parallel Computing

The most computationally expensive part of the GA calculations is the evaluation of the objective function. Fortunately, calculating objective functions in a GA is known as "an embarrassingly parallel problem" because it involves solving many similar, but independent tasks simultaneously in parallel, with little or no need for intertask coordination and communication. Consequently, it is possible to speed up the GA by outsourcing objective function calculations to local computer clusters or computing clouds. An affordable and interesting option is to use general purpose computing on graphical processing units (GPGPUs), since GPUs in common modern desktop computer graphics cards are already optimised for parallelism.

The GA library we have developed already supports parallel processing, thus, we should be able to extend our framework to apply parallel computing with little or no modification. Note, however, that we currently do not take advantage of parallel processing, since the evaluation of the cost function is performed by the server-side CPT, and not the client-side GA. In future work, we may consider improving the server-side CPT implementation by parallel processing, for example by recoding in Haskell and employing GPGUs for parallel evaluation of objective functions.

Case Study

To test the ability of our software framework to perform optimised CautD of an offshore crane, we used a real knuckleboom crane as a nominal benchmark against which our optimised crane design could be compared. The nominal crane has been designed, sold, and delivered by Seonics AS to a company in Baku, Azerbaijan. The crane had a total delivery price of approximately 2.9 million EUR.

Two KPIs were chosen as components of an objective function to be optimised: (i) the maximum safe working load SWL_{max} and (ii) the total crane weight W . Whilst the total crane delivery price is of great concern, we do not currently have price estimates as a function of crane design implemented in the CPT. Nevertheless, the total weight can to some extent be used as a proxy for price, because price will be correlated to weight, and one wants to minimise both measures. Moreover, these cranes are installed on-board vessels and reduced crane weight allows a higher

deadweight tonnage (DWT). Hence, weight is important for both capital and operating expenditure.

The maximum SWL, on the other hand, is a measure of the maximum lifting capacity of the crane in the entire workspace. In a particular crane configuration, with the tip of the crane in a particular position of the workspace where SWL is maximum, all other configurations with the crane tip in surrounding positions will have a SWL less than or equal to the maximum SWL.

The goal of the crane optimisation was thus to maximise SWL_{\max} while simultaneously minimising W . To achieve this goal, we let the optimisation function be a fitness function f_1 given by

$$f_1 = \frac{SWL_{\max}}{W}$$

The evaluation of f_1 will increase when SWL_{\max} increases and/or W decreases, and vice versa.

There are several other possible choices for objective functions, including handling with MOO, some of which are studied in our accompanying paper (Hameed et al., 2016). Here, we are mainly concerned with showing proof-of-concept, namely that our complete software framework performs as intended. A simple case study and a single objective function suffices for this purpose.

As optimisation variables, we chose four design parameters that greatly affect both SWL_{\max} and W : (i) the boom length; (ii) the jib length; (iii) the maximum pressure of the boom cylinder; and (iv) the maximum pressure of the jib cylinder. The parameter values were constrained to a range with minimum and maximum limits. All other design parameters were identical to those of the nominal crane.

RESULTS

Using the GA with a population size (set of candidate design solutions) of 100 and running for 50 generations, giving a grand total of 5,000 evaluated designs, took 98.4 minutes, including transfer times between the AIPO client and CPT server. The best design solution found is presented in Table 1, and shows that compared with the nominal benchmark crane, the maximum SWL improved from 100.0 tonnes to 142.1 tonnes, or an improvement of 42.2%, while simultaneously, the weight of crane was reduced from 50.8 tonnes to 44.0 tonnes, or an improvement of 13.5%. The objective function evaluated at the optimised design parameters was improved by 64.3%.

A graphical representation of the load charts of the nominal crane and the optimised crane design is shown in Figure 5. From the diagram, it is clear that the optimised design has resulted in a crane with a much smaller workspace than the nominal crane but with much higher lifting capacity.

DISCUSSION

In this paper, we have presented our first version of a generic and modular software framework for intelligent computer-automated product design. To demonstrate

proof-of-concept of the fully functioning complete system, we did a case study where our CautoD solution was able to improve the existing design of a real and delivered 50-tonnes, 2.9 million EUR knuckleboom crane with respect to the objectives of maximising the SWL and minimising the total crane weight. Whilst the results are sufficient for the purpose of proof-of-concept, we wish to emphasise that for real crane design, much more sophisticated objective functions must be constructed, able to handling a variety of different, possibly competing, design criteria. For example, our optimised design solution in the case study above outperforms the real crane on the objectives we have chosen but may fail on other design criteria not encompassed by our choice of objective function. For instance, the workspace of the optimised crane is much smaller compared to the real crane (see Fig. 5) and may not satisfy the needs of the company buying the crane.

Generic and Modular Software Architecture

We have implemented several software modules (see Figure 3) that together constitute the framework: (a) the server-side CPT calculates a fully specified crane design and a number of key performance indicators based on a set of about 120 input design parameters; (b) the client-side web GUI facilitates manually setting the design parameters of the CPT as well as providing an immediate visualisation of the resulting crane and its 2D workspace SWL load chart; (c) the client-side AIPO module uses a GA library for optimising the design parameters to achieve a crane design with desired performance. Communication between clients and server is achieved by means of the HTTP and WS protocols, and with standardised JSON messages as the data format.

Our framework being generic and modular, both client-side and server-side modules can easily be extended or replaced. On the client-side, crane designers may choose to create their own custom-made computer-automated design modules or GUIs that can communicate with the CPT to obtain desired crane designs and visualisations. We demonstrate the feasibility of this concept in an accompanying paper submitted concurrently (Hameed et al., 2016), in which we create a simple product optimisation client in Matlab that connects to the CPT and optimises various crane designs by means of a GA. On the server-side, crane designers can provide other product prototyping tools to which our existing AIPO module can connect and optimise the product design.

Furthermore, our research team is currently developing a winch prototyping tool to which our existing AIPO module can connect and optimise winch designs with little or no modification. This work will be published in the near future.

Complexity and Sensitivity

We have not analysed algorithm complexity nor sensitivity with respect to the design parameters. A core of $N = 120$ design parameters has been implemented in the crane calculator but only $n = 4 \ll N$ parameters were

measure	units	nominal	limits (min, max)	optimised	difference	change
boom length	mm	15,800	(12,000, 26,000)	12038	-3,762	-23.8%
jib length	mm	10,300	(6,000, 16,000)	6124	-4,176	-40.5%
boom cylinder max pressure	bar	315	(100, 400)	383	68	21.6%
jib cylinder max pressure	bar	215	(50, 300)	262	47	21.8%
SWL _{max}	kg	99,978	-	142,138	42,160	42.2%
W	kg	50,856	-	44,014	-6,842	-13.5%
$f_1 = \text{SWL}_{\text{max}}/W$	-	1.97	-	3.23	1.26	64.3%

Table 1: Optimised crane design compared with a nominal benchmark crane.

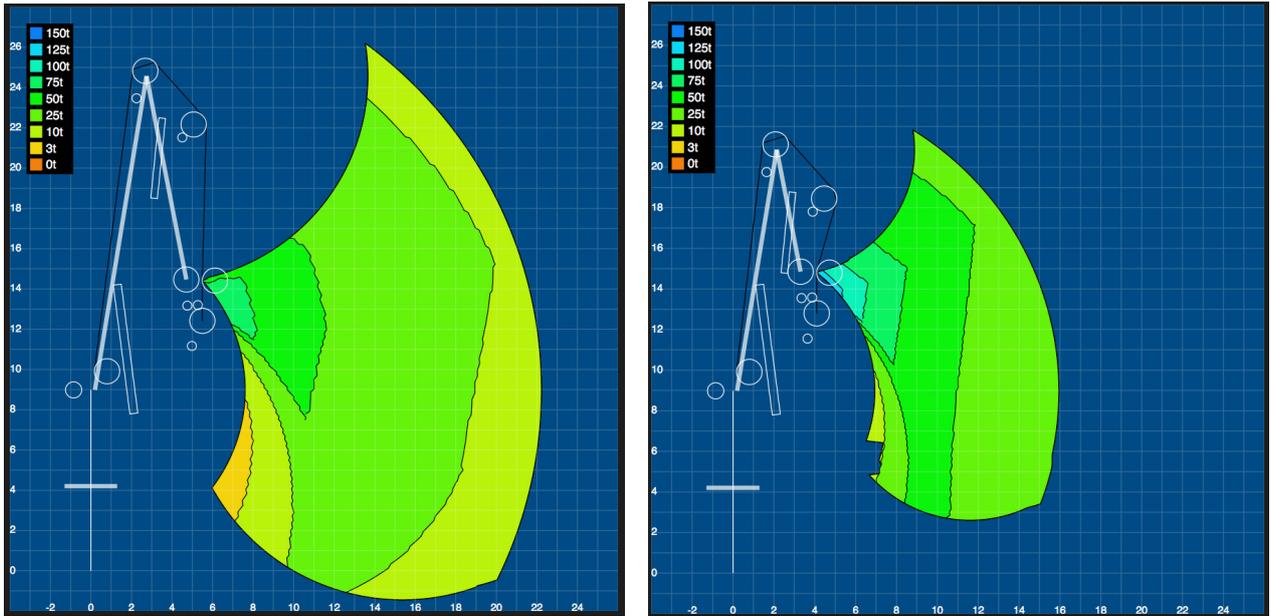


Figure 5: 2D SWL load charts for a nominal benchmark crane (left) and the optimised crane (right).

optimised by the GA in our case study. Now that we have demonstrated proof-of-concept of our framework with 4 parameters, we are in the process of investigating both the effects of varying both the choice of parameters and their values (sensitivity) and expanding our solution and let the number n grow (complexity). As the number of design parameters grows, the search space grows exponentially. Hence, parallel computing may be needed in order to find optimal solutions within reasonable time.

Intellectual Property Protection and Licensing

In its current implementation, our software framework partially supports protection of intellectual property as required by our industrial partners. Specifically, product designers who wants to use the web GUI to design offshore cranes will have no knowledge of the inner workings of the black box server-side CPT. Designers can also write their own customised client software to connect to the CPT, such as the Matlab crane optimisation client we have developed concurrently (Hameed et al., 2016).

The CautoD solution, on the other hand, consists of the AIPO module and its GA library and does not have a GUI frontend yet, which would have enabled black box usage. Thus, in its current state of implementation, users must

modify the code directly to obtain crane design results, and intellectual property is therefore not protected. We describe the possibility of extending the AIPO module with a web GUI in the next section.

For licensing, we have currently developed a username/password-protected version of the web GUI, which uses the WS/JSON interface. This can easily be extended to the HTTP/JSON interface as well. Using usernames/passwords enables the owner of the software to license usage of the CPT server on a time-limited basis. We may implement licensing for the CautoD solution, as well as more advanced licensing mechanisms, in future work.

AIPO GUI and Added Functionality

Product designers are experts in their own professional domains but will often lack prior knowledge of AI or programming. It is therefore essential to enable these domain experts to use our software framework while having to acquire as little as possible of this knowledge. Accordingly, we plan to develop a GUI for configuration of the AIPO module, possibly absorbing the functionality of the existing web GUI in Figure 3. This has several advantages. First, the code and intellectual property of the

AIPO module and related AI algorithms can be hidden and used as a black box. Second, the GUI could be used for configuration purposes, such as defining and using new objective functions or setting parameters or weights of a set of predefined objective functions. Moreover, such a GUI should enable the user to choose which AI algorithm should be used for optimisation and set its parameters. Third, the GUI could be used for visualisation, not only of the load chart, but of other relevant design aspects, thus enabling the product designer to quickly obtain an overview of the proficiency of the design and possible tradeoffs. Fourth, the GUI could let the user investigate existing product designs stored in a library, and possibly use these as a starting point for optimisation. Finally, the GUI can provide an intuitive and user-friendly method for doing all of the above when compared with text-based alternatives and direct programming of the software.

ACKNOWLEDGEMENTS

The SoftICE lab at NTNU in Ålesund wishes to thank ICD Software AS for their contribution towards the implementation of the simulator, and Seonics AS for providing documentation and insight into the design and manufacturing process of offshore cranes. We are also grateful for the support provided by Regionalt Forskningsfond (RFF) Midt-Norge and the Research Council of Norway through the VRI research projects *Artificial Intelligence for Crane Design (Kunstig intelligens for krandesign (KIK))*, grant no. 241238 and *Artificial Intelligence for Winch Design (Kunstig intelligens for vinsjdesign (KIV))*, grant no. 249171.

REFERENCES

- Alaliyat, S., Yndestad, H. and Sanfilippo, F. (2014), Optimisation of Boids Swarm Model Based on Genetic Algorithm and Particle Swarm Optimisation Algorithm (Comparative Study), Proceedings of the 28th European Conference on Modelling and Simulation.
- Arora, J. (2012), *Introduction to optimum design*, 3rd edn, Academic Press.
- Bak, M., Hansen, M. and Nordhammer, P. (2011), Virtual prototyping-model of offshore knuckle boom crane, Proceedings of the 24th International Congress on Condition Monitoring and Diagnostics Engineering Management, pp. 1242–1252.
- Bak, M. K. and Hansen, M. R. (2013), Analysis of Offshore Knuckle Boom Crane - Part One: Modeling and Parameter Identification, *Modeling, Identification and Control* 34(4), 157–174.
- 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., Osen, O. L. and Pedersen, B. S. (2015), A computer-automated design tool for intelligent virtual prototyping of offshore cranes, Proceedings of the 29th European Conference on Modelling and Simulation (ECMS'15), pp. 147–156.
- Bye, R. T. and Schaathun, H. G. (2014), An improved receding horizon genetic algorithm for the tug fleet optimisation problem, Proceedings of the 28th European Conference on Modelling and Simulation, pp. 682–690.
- Bye, R. T. and Schaathun, H. G. (2015), Evaluation heuristics for tug fleet optimisation algorithms: A computational simulation study of a receding horizon genetic algorithm, Proceedings of the 4th International Conference on Operations Research and Enterprise Systems (ICORES'15), pp. 270–282.
- Chan, S., Wong, M. and Ng, V. (1999), Collaborative solid modeling on the WWW, Proceedings of the 1999 ACM symposium on Applied computing, ACM, pp. 598–602.
- Chang, H.-C., Lu, W. F. and Liu, X. F. (1999), WWW-based collaborative system for integrated design and manufacturing, *Concurrent Engineering* 7(4), 319–334.
- Chaves, O., Nickelsen, M. and Gaspar, H. (2015), Enhancing Virtual Prototype in Ship Design Using Modular Techniques, *Proceedings 29th ECMS*.
- Choi, S. and Cheung, H. (2008), A versatile virtual prototyping system for rapid product development, *Computers in Industry* 59(5), 477–488.
- Chu, Y., Aesøy, V., Zhang, H. and Bunes, Ø. (2014), Modelling and simulation of an offshore hydraulic crane, 28th European Conference on Modelling and Simulation.
- Chu, Y., Hatledal, L. I., Sanfilippo, F., Asoy, V., Zhang, H. and Schaathun, H. G. (2015), Virtual prototyping system for maritime crane design and operation based on functional mock-up interface, OCEANS 2015-Genova, IEEE, pp. 1–4.
- De Sa, A. G. and Zachmann, G. (1999), Virtual reality as a tool for verification of assembly and maintenance processes, *Computers & Graphics* 23(3), 389–403.
- Erbatur, F., Hasançebi, O., Tütüncü, I. and Kılıç, H. (2000), Optimal design of planar and space structures with genetic algorithms, *Computers & Structures* 75(2), 209–224.
- Goldberg, D. E. (1983), Computer-aided gas pipeline operation using genetic algorithms and rule learning, PhD thesis, University of Michigan Ann Arbor.
- Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Professional.
- Hameed, I. A., Osen, O. L., Bye, R. T., Pedersen, B. S. and Schaathun, H. G. (2016), Intelligent computer-automated crane design using an online crane prototyping tool, Proceedings of the 30th European Conference on Modelling and Simulation (ECMS'16) (submitted for publication).
- Hare, W., Nutini, J. and Tesfamariam, S. (2013), A survey of non-gradient optimization methods in structural engineering, *Advances in Engineering Software* 59, 19–28.
- Hatledal, L. I., Sanfilippo, F., Chu, Y. and Zhang, H. (2015), A voxel-based numerical method for computing and visualising the workspace of offshore cranes, ASME 2015 34th International Conference on Ocean, Offshore and Arctic Engineering, American Society of Mechanical Engineers, pp. V001T01A012–V001T01A012.
- Haupt, R. L. and Haupt, S. E. (2004), *Practical Genetic Algorithms*, 2nd edn, Wiley.
- He, B., Tang, W. and Cao, J. (2014), Virtual prototyping-based multibody systems dynamics analysis of offshore crane, *The International Journal of Advanced Manufacturing Technology* 75(1-4), 161–180.
- Holland, J. H. (1975), *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*, University of Michigan Press, Oxford, England.
- Kamentsky, L. and Liu, C. (1963), Computer-automated design of multiform print recognition logic, *IBM Journal of Research and Development* 7(1), 2–13.
- Kaveh, A. and Talatahari, S. (2009), Hybrid algorithm of harmony search, particle swarm and ant colony for structural design optimization, *Harmony search algorithms for structural design optimization*, Springer, pp. 159–198.
- Kim, H., Lee, J.-K., Park, J.-H., Park, B.-J. and Jang, D.-S. (2002), Applying digital manufacturing technology to ship production and the maritime environment, *Integrated Manufacturing Systems* 13(5), 295–305.
- Kirkpatrick, S. (1984), Optimization by simulated annealing: Quantitative studies, *Journal of statistical physics* 34(5-

- 6), 975–986.
- Mujber, T., Szecsi, T. and Hashmi, M. (2004), Virtual reality applications in manufacturing process simulation, *Journal of materials processing technology* 155, 1834–1838.
- Narayan, K. L., Rao, K. M. and Sarcar, M. (2008), *Computer aided design and manufacturing*, PHI Learning Pvt. Ltd.
- Park, H.-S. and Le, N.-T. (2012), Modeling and controlling the mobile harbour crane system with virtual prototyping technology, *International Journal of Control, Automation and Systems* 10(6), 1204–1214.
- Pawlus, W., Choux, M., Hovland, G., Hansen, M. R. and Øydn, S. (2014), Modeling and simulation of an offshore pipe handling machine, Proceedings of the 55th Conference on Simulation and Modelling (SIMS 55), pp. 277–284.
- Peng, Y., Yancong, L., Yongjun, S. and Guande, L. (2010), Optimum Piping Design on Offshore Platform Based on Improved Adaptive Genetic Algorithm, Proceedings of the 2010 WASE International Conference on Information Engineering-Volume 01, IEEE Computer Society, pp. 50–53.
- Pezeshk, S., Camp, C. and Chen, D. (2000), Design of nonlinear framed structures using genetic optimization, *Journal of Structural Engineering* 126(3), 382–388.
- Pratt, M. J. (1995), Virtual prototypes and product models in mechanical engineering, *Virtual Prototyping–Virtual environments and the product design process* 10, 113–128.
- Rajeev, S. and Krishnamoorthy, C. (1992), Discrete optimization of structures using genetic algorithms, *Journal of Structural Engineering* 118(5), 1233–1250.
- Sanfilippo, F., Hatledal, L. I., Schaathun, H. G., Pettersen, K. Y. and Zhang, H. (2013), A universal control architecture for maritime cranes and robots using genetic algorithms as a possible mapping approach, Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on, IEEE, pp. 322–327.
- Sanfilippo, F., Hatledal, L., Zhang, H., Rekdalsbakken, W. and Pettersen, K. (2015), A wave simulator and active heave compensation framework for demanding offshore crane operations, Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference on, IEEE, pp. 1588–1593.
- Sarshar, M., Christiansson, P. and Winter, J. (2004), Towards virtual prototyping in the construction industry: the case study of the DIVERCITY project, Proceedings of the World IT Conference for Design and Construction, pp. 18–24.
- Schöning, C. (2014), Virtual prototyping and optimisation of microwave ignition devices for the internal combustion engine, PhD thesis, University of Glasgow.
- Schoning, L.-C. and Li, Y. (2013), Multivariable optimisation of a Homogeneous Charge Microwave Ignition system, Automation and Computing (ICAC), 2013 19th International Conference on, IEEE, pp. 1–5.
- Shen, Q., Gausemeier, J., Bauch, J. and Radkowski, R. (2005), A cooperative virtual prototyping system for mechatronic solution elements based assembly, *Advanced Engineering Informatics* 19(2), 169–177.
- Shyamsundar, N. and Gadh, R. (2002), Collaborative virtual prototyping of product assemblies over the Internet, *Computer-Aided Design* 34(10), 755–768.
- Černý, V. (1985), Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm, *Journal of optimization theory and applications* 45(1), 41–51.
- Waly, A. F. and Thabet, W. Y. (2003), A virtual construction environment for preconstruction planning, *Automation in construction* 12(2), 139–154.
- Weyrich, M. and Drews, P. (1999), An interactive environment for virtual manufacturing: the virtual workbench, *Computers in industry* 38(1), 5–15.
- Wöhlke, G. and Schiller, E. (2005), Digital planning validation in automotive industry, *Computers in industry* 56(4), 393–405.
- Xu, X. and Luo, Y. (2010), Force finding of tensegrity systems using simulated annealing algorithm, *Journal of structural engineering* 136(8), 1027–1031.
- Yerrapathruni, S. (2003), Using 4 D CAD and Immersive Virtual Environments to Improve Construction Planning, PhD thesis, Architectural Engineering.
- Zhang, J., Zhan, Z., Lin, Y., Chen, N., Gong, Y.-j., Zhong, J.-h., Chung, H., Li, Y. and Shi, Y. (2011), Evolutionary computation meets machine learning: A survey, *Computational Intelligence Magazine, IEEE* 6(4), 68–75.

AUTHOR BIOGRAPHIES

ROBIN T. BYE⁶ 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 NTNU in Ålesund (formerly Aalesund University College) as a researcher in 2008 and has since 2010 been an associate professor in automation engineering. His research interests belong to the fields of artificial intelligence, cybernetics, and neuroengineering.

OTTAR L. OSEN is MSc in Cybernetics from the Norwegian Institute of Technology in 1991. He is the head of R&D at ICD Software AS and an assistant professor at NTNU in Ålesund.

BIRGER SKOGENG PEDERSEN graduated from NTNU in Ålesund with a BE in automation engineering and is a former employee at ICD Software AS, during which time he worked on the research projects described in this paper. He is currently a MSc student of simulation and visualisation as well as a employed as a project manager in maritime technology at NTNU in Ålesund.

IBRAHIM A. HAMEED has a BSc and a MSc in Industrial Electronics and Control Engineering, Menofia University, Egypt, a PhD in Industrial Systems and Information Engineering from Korea University, S. Korea, and a PhD in Mechanical Engineering, Aarhus University, Denmark. He has been working as an associate Professor at NTNU in Ålesund since 2015. His research interests includes artificial Intelligence, optimization, control systems and robotics.

HANS GEORG SCHAATHUN Hans Georg Schaathun graduated from the University of Bergen with cand.mag. in 1996 (Mathematics, Economics, and Informatics), cand.scient. in 1999 (Industrial and Applied Mathematics and Informatics), and dr.scient. in 2002 (Informatics – Coding Theory), all from the University of Bergen, Norway. He was a lecturer in coding and cryptography at the University of Bergen 2002 and a postdoc in 2003-2006. As a lecturer and senior lecturer in computer science at the University of Surrey, England 2006-2010, his research focused on multimedia security including applications of coding theory and steganalysis using machine learning. He joined Aalesund University College (now NTNU in Ålesund) and became a professor in 2011. His current research focus is software engineering and pedagogy.

⁶www.robinbye.com

On Usage of EEG Brain Control for Rehabilitation of Stroke Patients

Tom Verplaetse*, Filippo Sanfilippo†, Adrian Rutle‡, Ottar L. Osen§, and Robin T. Bye§

* Faculty of Engineering and Architecture

Department of Industrial Technology and Construction
Ghent University

Valentin Vaerwyckweg 1, BE-9000 Gent, Belgium

Email: tom.verplaetse@ugent.be

† Department of Engineering Cybernetics

Norwegian University of Science and Technology
NO-7491 Trondheim, Norway

Email: filippo.sanfilippo@ntnu.no

‡ Department of Computing, Mathematics and Physics
Bergen University College

P.O. Box 7030, NO-5020 Bergen, Norway

Email: adrian.rutle@hib.no

§ Software and Intelligent Control Engineering Laboratory
Faculty of Engineering and Natural Sciences

Norwegian University of Science and Technology
NTNU in Ålesund, Postboks 1517, NO-6025 Ålesund, Norway

Email: {robin.t.bye, ottar.l.osen}@ntnu.no

KEYWORDS

Motor imagery, brain-computer interface, virtual reality, game-stimulated rehabilitation, low-cost commercial-off-the-shelf products.

ABSTRACT

This paper demonstrates rapid prototyping of a stroke rehabilitation system consisting of an interactive 3D virtual reality computer game environment interfaced with an EEG headset for control and interaction using brain waves. The system is intended for training and rehabilitation of partially monoplegic stroke patients and uses low-cost commercial-off-the-shelf products like the Emotiv EPOC EEG headset and the Unity 3D game engine. A number of rehabilitation methods exist that can improve motor control and function of the paretic upper limb in stroke survivors. Unfortunately, most of these methods are commonly characterised by a number of drawbacks that can limit intensive treatment, including being repetitive, uninspiring, and labour intensive; requiring one-on-one manual interaction and assistance from a therapist, often for several weeks; and involve equipment and systems that are complex and expensive and cannot be used at home but only in hospitals and institutions by trained personnel. Inspired by the principles of mirror therapy and game-stimulated rehabilitation, we have developed a first prototype of a game-like computer application that tries to avoid these drawbacks. For rehabilitation purposes, we deprive the patient of the view of the paretic hand

while being challenged with controlling a virtual hand in a simulated 3D game environment only by means of EEG brain waves interfaced with the computer. Whilst our system is only a first prototype, we hypothesise that by iteratively improving its design through refinements and tuning based on input from domain experts and testing on real patients, the system can be tailored for being used together with a conventional rehabilitation programme to improve patients' ability to move the paretic limb much in the same vein as mirror therapy. Our proposed system has several advantages, including being game-based, customisable, adaptive, and extendable. In addition, when compared with conventional rehabilitation methods, our system is extremely low-cost and flexible, in particular because patients can use it in the comfort of their homes, with little or no need for professional human assistance. Preliminary tests are carried out to highlight the potential of the proposed rehabilitation system, however, in order to measure its efficiency in rehabilitation, the system must first be improved and then run through an extensive field test with a sufficiently large group of patients and compared with a control group.

INTRODUCTION

A stroke is a medical condition when blood supply to the brain is interrupted or reduced, usually because a blood vessel bursts or is blocked by a clot (*Stroke, Cerebrovascular Accident* | World Health Organization, 2016). As a consequence, brain cells are deprived of oxygen and begin to die, resulting in a loss of control of muscles that are controlled by the dead area of the brain. The effects of a

Corresponding author: Robin T. Bye.

stroke may vary from minor problems, such as temporary weakness of an arm or leg, to permanent paralysis on one side of the body or losing the ability to speak, depending on which part of the brain is injured and how severely it is affected. Some people recover completely from strokes, but more than 2/3 of survivors will have some type of disability (*What is Stroke?* | National Stroke Association, 2016). A typical result of stroke is monoplegia, where the patient loses control over a single limb, usually an arm.

The focus of this paper is on the development of a first prototype of a rehabilitation system for partially monoplegic patients with a paretic hand. Such patients have not completely lost all motor function of the limb and therefore have the potential for rehabilitation. The main components of our system are a computer application with a game-like 3D virtual environment, and an electroencephalography (EEG) brain-computer interface (BCI) providing control inputs to the application by means of brain waves.

In the following sections, we first provide some background on existing stroke rehabilitation methods, with a particular focus on mirror therapy and game-stimulated rehabilitation; EEG and BCI; the steady state visually evoked potential (SSVEP); and our motivation and aim. We proceed with presenting our proposed method, before turning our attention to the implementation details of our system. Finally, we present some preliminary test results and a discussion of our work.

BACKGROUND

Stroke Rehabilitation

Whilst stroke rehabilitation has come a long way since the early ages of medicine, it is still an active field of research with room for improved methodologies. Before the 1950s, stroke rehabilitation was practically non-existent and damage control was the only approach. Thomas Twitchell was one of the first to accurately describe the possible recovery after a stroke (Twitchell, 1951). He posed that it was possible to achieve full recovery within a certain period of time if stroke patients underwent a suitable rehabilitation programme. Around the same time Signe Brunnstrom developed the Brunnstrom approach (Brunnstrom, 1966; Brunnström, 1970), an approach to determine the stage of the recovery and consequently evaluate different rehabilitation techniques.

As noted by Nudo and Duncan (2004), an increasing number of studies demonstrate the property of neuroplasticity in the brain, with sensorimotor regions of the brain undergoing both structural and functional alterations as a function of use and injury. The development of interventions for stroke rehabilitation have been shown to result not only in adaptive reorganization in the cerebral cortex but may also invigorate recovery in the impaired limb months or years after the stroke (Nudo and Duncan, 2004). A number of various rehabilitation methods and their effects in improving upper-extremity motor control and functioning have been reported in the literature, including

exercise training of the paretic arm (Kwakkel, Wagenaar, Twisk, Lankhorst and Koetsier, 1999), impairment-oriented training of the arm (Platz, Eickhof, Van Kaick, Engel, Pinkowski, Kalok and Pause, 2005), functional electric stimulation (Ring and Rosenthal, 2005), robotic-assisted rehabilitation (Masiero, Celia, Rosati and Armani, 2007), and bilateral arm training (Summers, Kagerer, Garry, Hiraga, Loftus and Cauraugh, 2007). However, as pointed out by Yavuzer, Selles, Sezer, Sütbeyaz, Bussmann, Köseoğlu, Atay and Stam (2008), most of the rehabilitation methods that exist for the paretic upper extremity are labour intensive, and require personal interaction and instructions from trained personnel such as therapists for several weeks, which makes it difficult to ensure proper intensive treatment for all patients. In addition, the equipment and systems used in stroke rehabilitation are often expensive, non-portable, and complex, thus requiring being located in a hospital or institution and operated by trained medical personnel.

Mirror Therapy

Contrary to most rehabilitation methods, mirror therapy is a simple, inexpensive, and patient-directed rehabilitation method that has been shown to improve hand functioning when used in conjunction with conventional stroke rehabilitation programmes (e.g., Yavuzer et al., 2008). Yavuzer and colleagues instructed patients to perform wrist and finger extension and flexion movements simultaneously with both their paretic and nonparetic hand, whilst the paretic hand was hidden from sight. While doing the movements, patients watched a mirror image of their normal functioning hand, thus tricking the brain into believing that the paretic hand was actually able to perform the movements. Compared with a control group of patients who received sham treatment, the patients who received mirror treatment significantly improved their motor recovery as measured by Brunnstrom stages.

The findings of Yavuzer et al. (2008) have later been enforced in a metastudy by Thieme, Mehrholz, Pohl, Behrens and Dohle (2013), who examined 14 studies that compared mirror therapy with other interventions, and found that compared with all the other interventions, mirror therapy had a significant effect on motor function, although effects were dependent on the type of control intervention. Thieme and colleagues also found that mirror therapy was found to significantly improve everyday living activities and reduce pain but found limited evidence for improving visuospatial neglect.

Game-Stimulated Stroke Rehabilitation

There are several studies in the literature of using serious games for game-stimulated stroke rehabilitation (e.g., Burke, McNeill, Charles, Morrow, Crosbie and McDonough, 2009a,b; Alankus, Lazar, May and Kelleher, 2010; Yavuzer, Senel, Atay and Stam, 2008; Vogiatzaki and Krukowski, 2014; Lewis, Woods, Rosie and McPherson, 2011; Joo, Yin, Xu, Thia, Chia, Kuah and He, 2010). Effective stroke rehabilitation requires intensive

and repetitive exercises that can be demotivating for the patient, however, game-stimulated rehabilitation can improve patient motivation, enjoyment, and engagement (e.g., Yavuzer, Senel, Atay and Stam, 2008; Lewis et al., 2011; Joo et al., 2010).

For example, Joo et al. (2010) found that using the Nintendo Wii as an adjunct to conventional rehabilitation of patients with post-stroke upper limb weakness was more enjoyable than conventional therapy and showed that there were small but statistically significant improvements in the Fugl-Meyer Assessment and Motricity Index scores.

Another example is that of Yavuzer, Senel, Atay and Stam (2008), who examined the effects of using the PlayStation EyeToy Games on upper extremity motor recovery and upper extremity-related motor functioning of patients with subacute stroke. They found that the functional independence measure (FIM) significantly improved in the patients with the EyeToy intervention compared to the control group. However, no significant differences were found between the groups for the Brunnstrom stages for hand and upper extremity.

Another advantage of using serious games is that the therapy can take place in the patient's home, making it easier for the patient to complete the necessary number of exercise repetitions in her own time, and with the game easily customized to the patient's needs and progression (Alankus et al., 2010).

In addition to the use of game development platforms such as the PlayStation and Nintendo Wii mentioned above, the Unity 3D game engine has also been used, as have control interfaces such as Microsoft Kinect, CyberGlove, Rutgers RMII Master, Leap Motion, Emotiv EPOC EEG, the Viacon camera system, and electromyography (EMG) measurements, and 3D projectors, thus enabling a great variety of game-stimulated stroke rehabilitation methods (e.g., see Vogiatzaki and Krukowski, 2014).

Electroencephalography (EEG)

EEG is an electrophysiological monitoring method that measures the natural electric potential on the scalp (Niedermeyer and da Silva, 2005). Physiologically, EEG power reflects the number of neurons that discharge synchronously (Klimesch, 1999). This electric potential is a result of brain activity and behaves in a periodic, wavelike fashion referred to as brain waves and can be recorded with a portable EEG headset such as the Emotiv EPOC EEG (e.g., Duvinage, Castermans, Petieau, Hoellinger, Cheron and Dutoit, 2013). The brain waves are divided into frequency bands, where each band corresponds to different brain functions. The EEG frequency bands are categorised as the delta (< 4 Hz), theta (4–7 Hz), alpha (8–15 Hz), beta (16–31 Hz), and gamma (> 32 Hz) bands, of which the alpha band, which is active during an alert and cognitive state of the patient (Klimesch, 1999), and the beta band, which is closely related to purposive movement (Niedermeyer and da Silva, 2005), are most important with respect to stroke rehabilitation.

Brain waves and brain wave pattern recognition have been widely investigated in the recent literature. For instance, this technology has been employed for monitoring and prevention purposes, motivated by the fact that there is physiologic coupling of EEG morphology, frequencies, and amplitudes with cerebral blood flow (Ueki, Linn and Hossmann, 1988). Intraoperative continuous electroencephalographic monitoring (CEEG) is an established modality that is used to detect cerebral ischemia during carotid surgery. These facts have generated interest in applying EEG/CEEG in the intensive care unit to monitor cerebral ischemia. There is also evidence that EEG and CEEG add value to early diagnosis, outcome prediction, patient selection for treatment, clinical management, and seizure detection in acute ischemic stroke (AIS) (e.g., Jordan, 2004).

Motor Imagery Brain-Computer Interface (MI-BCI)

In a large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain-computer interface (MI-BCI) presented by Ang, Guan, Chua, Ang, Kuah, Wang, Phua, Chin and Zhang (2011), it was shown how BCI technology has the prospects of helping stroke survivors by enabling the interaction with their environment through brain signals rather than through muscles, and restoring motor function by inducing activity-dependent brain plasticity. The same work presented a clinical study on the extent of detectable brain signals from a large group of 54 stroke patients in using an EEG-based MI-BCI. A clinical study that investigated the ability of hemiparetic stroke patients in operating an EEG-based MI-BCI was also presented in (Ang, Guan, Chua, Ang, Kuah, Wang, Phua, Chin and Zhang, 2010). This work also assessed the efficacy in motor improvements on the stroke-affected upper limb using EEG-based MI-BCI with robotic feedback neuro-rehabilitation compared to robotic rehabilitation that delivers movement therapy.

Recent studies have found distinct cortical physiology associated with contralesional limb movements in regions distinct from primary motor cortex (e.g., see Fok, Schwartz, Wronkiewicz, Holmes, Zhang, Somers, Bundy and Leuthardt, 2011). These findings allow researchers to implement closed-loop interaction systems with valuable kinesthetic feedback for the user. For instance, based on these findings, a BCI that localises and acquires these brain signals to drive a powered hand orthotic was designed and implemented in (Fok et al., 2011). In this work, the patient's hand was guided with appropriate force feedback, thus enabling the hand to open and close.

Steady State Visually Evoked Potential (SSVEP)

Evoked potentials are specific patterns in brain activity which are caused by inputs to the patient from the inside or the outside of the body. These potentials can be recorded through the use of EEG. Steady state potentials are recorded potentials which show a phase, frequency and amplitude that are directly related to the input that caused the potential. If the patient is exposed to an input method

which delivers the input at a steady frequency and brain activity is recorded at the same or a related frequency, this brain activity is called a steady state evoked potential. If for example the input in question is a flashing screen placed in front of the patient, it is called a steady state visually evoked potential (SSVEP) (Misulis, Fakhoury and Spehlmann, 2001), and may be used to increase brain activity in certain desired EEG bands.

According to a survey by Zhu, Bieger, Molina and Aarts (2010), BCI systems based on the SSVEP provide a higher level of information throughput and require shorter training than BCI systems using that are not augmented with SSVEP. On the negative side, repetitive visual stimuli modulated at certain frequencies can provoke epileptic seizures or induce fatigue (Fisher, Harding, Erba, Barkley and Wilkins, 2005).

With the increased activation of selected EEG bands, we hypothesise that in line with the previously mentioned properties of neuroplasticity and adaptive reorganization of the cerebral cortex (Nudo and Duncan, 2004), employing SSVEP could result in a faster rehabilitation process.

Motivation and Aim

Most of the stroke rehabilitation methodologies described in previous sections require technical assistance to be provided to the patient by professional and skilled personnel. In addition, these methods typically are dependent on systems and equipment that are normally very costly and therefore only available at specialised medical centres. Hence, patients cannot perform their rehabilitation programmes at home but instead have to physically commute to reach these centres, with accompanying cost, time expenditure, and physical stress. However, as we describe above, stroke rehabilitation methods like mirror therapy and game-stimulated rehabilitation are able to counteract some of these drawbacks and improve motor functioning when used as an adjunct to conventional rehabilitation therapy.

Inspired by mirror therapy and game-stimulated rehabilitation, our aim is to combine the two and present a first prototype of a flexible and easy-to-use stroke rehabilitation system for the paretic hand consisting of an interactive 3D virtual reality computer game environment interfaced with an EEG headset for control and interaction using brain waves. Our proposed solution is a customisable and extensible low-cost framework that allows patients to perform their rehabilitation programme in the comfort of their house, with potentially no need for human assistance, adapting to patients' progression and skill, and that can be extended to other interfaces and external devices, e.g., a robotic exoskeleton for manipulating the paretic hand.

METHOD

The following sections provide details on the Emotiv EPOC EEG headset, including data acquisition, training of mental commands, 3D modelling of the paretic hand, and a description of the first prototype of a computer application containing the 3D virtual environment using the Unity 3D game engine.

EEG Data Acquisition

The Emotiv EPOC EEG headset is a high resolution, multi-channel, portable system that has been designed for practical research applications (Emotiv, 2016b). It has 14 EEG channels with its sensors placed according to the international 10-20 system¹ such that EEG activity from the following brain areas is measured (see Figure 1): AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. Each channel samples the small variations in electric potential on the scalp with a dynamic range of ± 4.17 mV, a resolution of $0.51\mu\text{V}$, and at a frequency of 2048 Hz, subsequently filtered and downsampled to 128 Hz (Emotiv, 2016a). The EEG signal data are transmitted wirelessly via bluetooth to a USB receiver that connects to a computer.

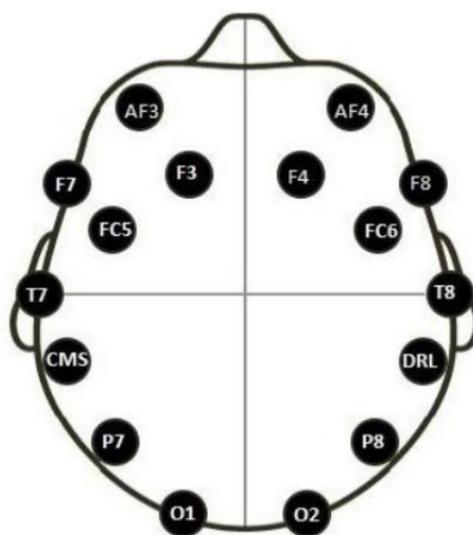


Figure 1: Positions of electrodes.

Training of Mental Commands

Emotiv software is able to perform pattern recognition of the received EEG data, thus building up a library of trained mental commands calibrated to each user of the equipment. Up to four different user-defined commands can be learnt and stored by the software. These commands typically have labels such as 'push', 'pull', 'lift', 'drop', 'left', 'right', however, by employing the application programming interface (API), one is free to access these commands and use them for whatever purpose the software developer finds suitable.

In our case, we used the Emotiv API to interface such commands to Unity for controlling a virtual hand in a 3D virtual environment but restricted to the 2D horizontal plane. In the reaching phase of a hand movement, the commands to be learnt by a user is therefore 'left' and 'right' (along the x-axis) and 'forward' and 'back' (along the y-axis), whereas in the grasping phase, the commands are 'open' and 'close'.

¹Wikipedia: [http://en.wikipedia.org/wiki/10-20_system_\(EEG\)](http://en.wikipedia.org/wiki/10-20_system_(EEG))

Whilst Emotiv provide their own software training environment (in the Emotiv Control Panel), we have conveniently integrated this environment in our own Unity application. This serves the purpose of making our application standalone, but more importantly, lets us design our own training environment tailored for stroke rehabilitation patients.

It is well known from the literature (e.g., Pfurtscheller, Flotzinger and Kalcher, 1993) that EEG signals will display characteristic changes just prior to making a physical movement. Indeed, Emotiv (2016b) make use of this property in their own software, whereby making a facial expression such a lifting an eyebrow or blinking an eye will elicit a particular EEG pattern that the software can easily recognise and convert to a mental command. Therefore, as for mirror therapy, we believe it is crucial that patients try to physically move both the paretic and nonparetic hand while performing EEG training. The purpose of this is to link the particular EEG patterns that emerge when patients move their hands with brain control of the virtual hand.

3D Model of the Paretic Hand

The success of mirror therapy is based on tricking the brain of the patient into believing that the paretic hand moves without any problem when in fact it is just the mirror image of the functioning nonparetic hand that the patient observes. Whilst a 3D virtual representation can never be made as perfect as a mirror image, we hypothesise that by using a virtual hand that is very realistic both in behaviour and in looks, coupled with direct brain control using EEG, we can get a positive effect in rehabilitation similar to that of mirror therapy.

For creating a realistic 3D model of the paretic hand, we used Blender,² which is a free and open source 3D creation suite. It supports a variety of 3D modelling aspects, including rigging, animation, simulation, rendering, compositing, and motion tracking. The resulting virtual hand is visually realistic and equipped with an internal set of finger joints that can be accessed and individually controlled by a computer programme. Importantly, however, to be realistic, the virtual joints must be programmed to move in a synergy, that is, a coordinated manner, just like real hand. After development, the 3D hand model was imported into Unity.

Unity 3D Application

The Unity 3D application is made up of four different scenes: a main menu, a settings panel, a training environment, and a game rehabilitation environment. Navigation between these scenes is done by using buttons that are readily available during the game. Fig. 2 shows some screenshots of scenes from the application, and how they are connected.

Main Menu: The main menu is the boot scene of the application, and is used to navigate to all the other scenes.

²www.blender.org

Settings Panel: The purpose of the settings panel is to calibrate the application to have the best possible effect on the patient. There are four different settings: (i) Hand movement speed, which is the speed with which the hand moves in an in the horizontal x-y plane. The faster the hand moves the more difficult it becomes for the patient to control the hand; (ii) The hand close speed, which is the time it takes for the hand to close and grasp target object (a ball, in this case) inside the environment. The longer the hand close speed, the longer the patient has to focus on the close command; (iii) target score, which is the number of objects (balls) the patient must successfully be able to grasp before the game ends; and (iv) SSVEP frequency, which is the frequency with which the screen's background colour will flash in order to evoke the SSVEP. The ideal frequency will have to be determined by physicians, likely through an experimental procedure, in order to achieve the optimal level of brain activation.

Training Environment: The training environment has been developed by Emotiv. It is basically the same training environment they use in their own native software but ported to the Unity game platform, thus enabling us to use it as a separate scene in our Unity application. The EEG headset can be set up in this scene and the different mental commands can be trained. Since EEG patterns vary across users, the software stores and maintains a calibrated profile for each user of the headset.

Game Rehabilitation Environment: To limit the initial work, we have only fully implemented a virtual game rehabilitation environment for the left paraplegic hand, and only for a single kind or exercise. The extension to the right hand is straight forward due to the dual properties of the left and right hand and intended for future work. The developed environment can also serve as a template if other limbs need to be considered.

In the game rehabilitation environment, the patient is presented with a top view of the virtual hand and a randomly positioned target object, which in the current implementation is a ball. The purpose of the game is for the patient to move the hand in the horizontal x-y plane. This is achieved by the patient simultaneously sending two mental movement commands (e.g., 'right' and 'forward' for a movement in the positive x and y directions).

When the target is approached, the game view zooms in to a close-up and the hand speed slows. The patient must then send a mental 'close' command in order to grasp the target object. The 'close' command must last for a minimum duration (can be adjusted in the settings menu) for the virtual hand to successfully grasp the object, otherwise, the hand will open and the patient will have to try again.

After successfully having grasped the object, the score counter is incremented and the game resets with the target in a new random position. The exercise is repeated until the desired number of successful reach-and-grasps have been reached (can be adjusted in the settings menu).

Importantly, for both the reaching phase and the grasping phase of the exercise, we hypothesise that the patient should also try to move his physical paretic hand and

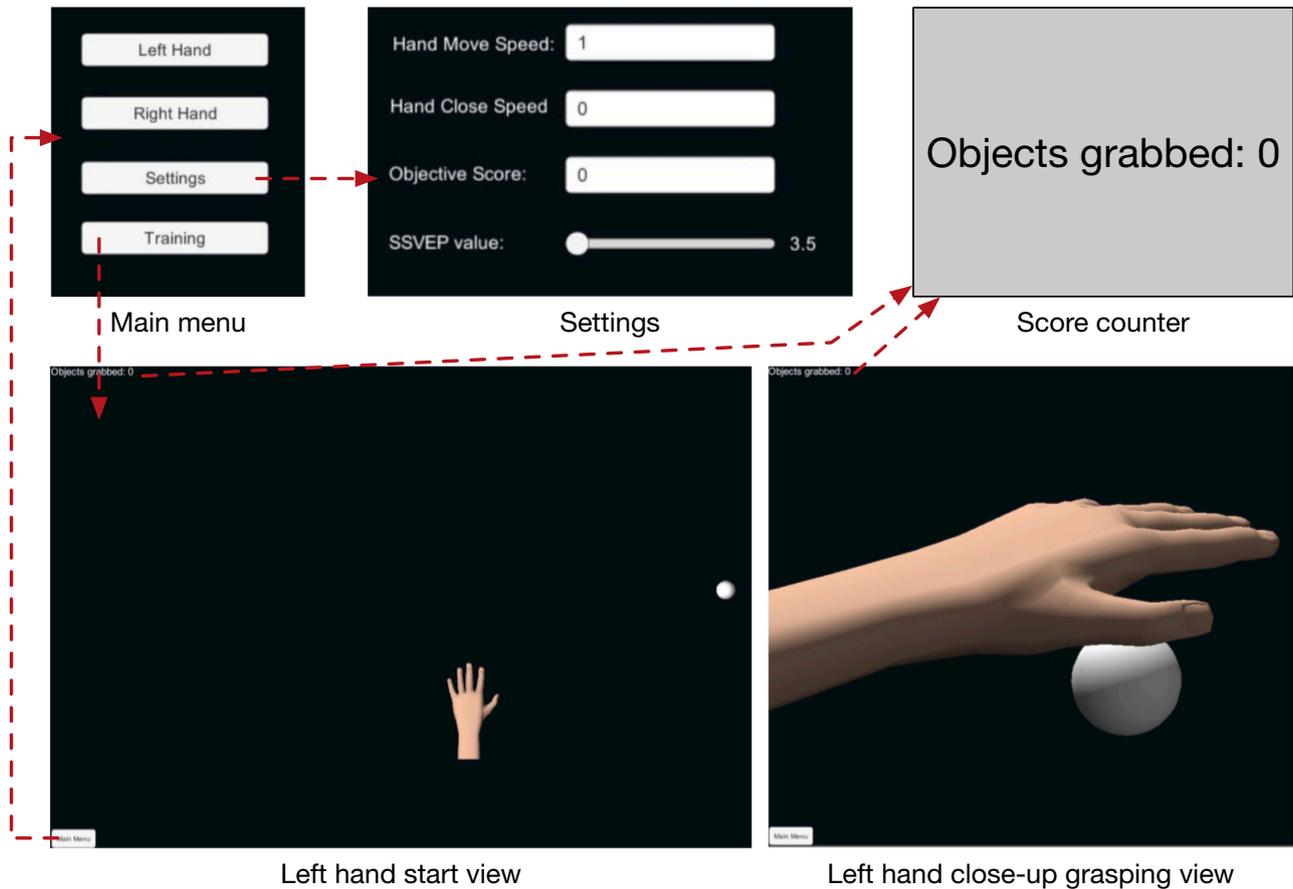


Figure 2: Selected scenes from the stroke rehabilitation application: main menu (top left); settings panel (top middle); and game rehabilitation environment, consisting of left hand start view (bottom left), left hand close-up grasping view (bottom right), and score counter (top right).

his well-functioning nonparetic hand, in order for the rehabilitation to get a positive effect similar to mirror therapy.

Finally, we highlight that during the rehabilitation exercise, the background can be set to be constantly flashing between two colours with the purpose of evoking the SSVEP for higher EEG activation and better mental control.

Application Overview and System Diagram

A high-level overview of the application is shown in Fig. 3, which shows how the application is built on top of the Emotiv EPOC EEG headset interfaced by an API with the Unity 3D game engine. Some of the features of the application are highlighted.

The system diagram in Fig. 4 shows the different system modules, data flow and usage modes of the stroke rehabilitation system. The top box in stapled blue line shows modules that provide the system setup. The SSVEP module is used for setting SSVEP parameters such as colours and flashing frequency. Blender is used to create 3D models such as a virtual hand. Data (settings and 3D model) is passed as inputs to the Unity 3D game environment. The game environment is presented to the user via the display. The user's EEG brain waves are sampled by the EEG headset and pattern-matched with

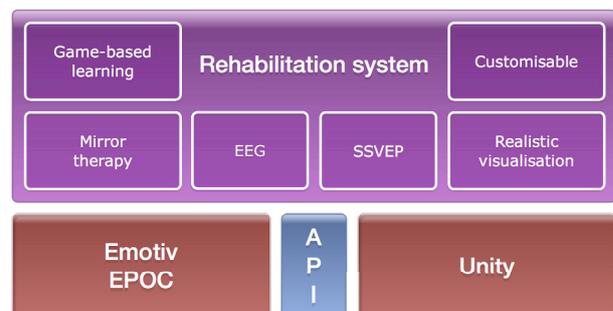


Figure 3: System overview.

a library of brain wave patterns and corresponding actions in order to generate the correct action, which is fed to back to Unity.

PRELIMINARY TEST RESULTS

Having partially monoplegic patients testing this first prototype of our stroke rehabilitation system would have been unethical, as the system needs further development in close cooperation with medical experts before being tested on real patients. Nevertheless, the first author, a healthy 22-year-old male at the time, constantly tested the

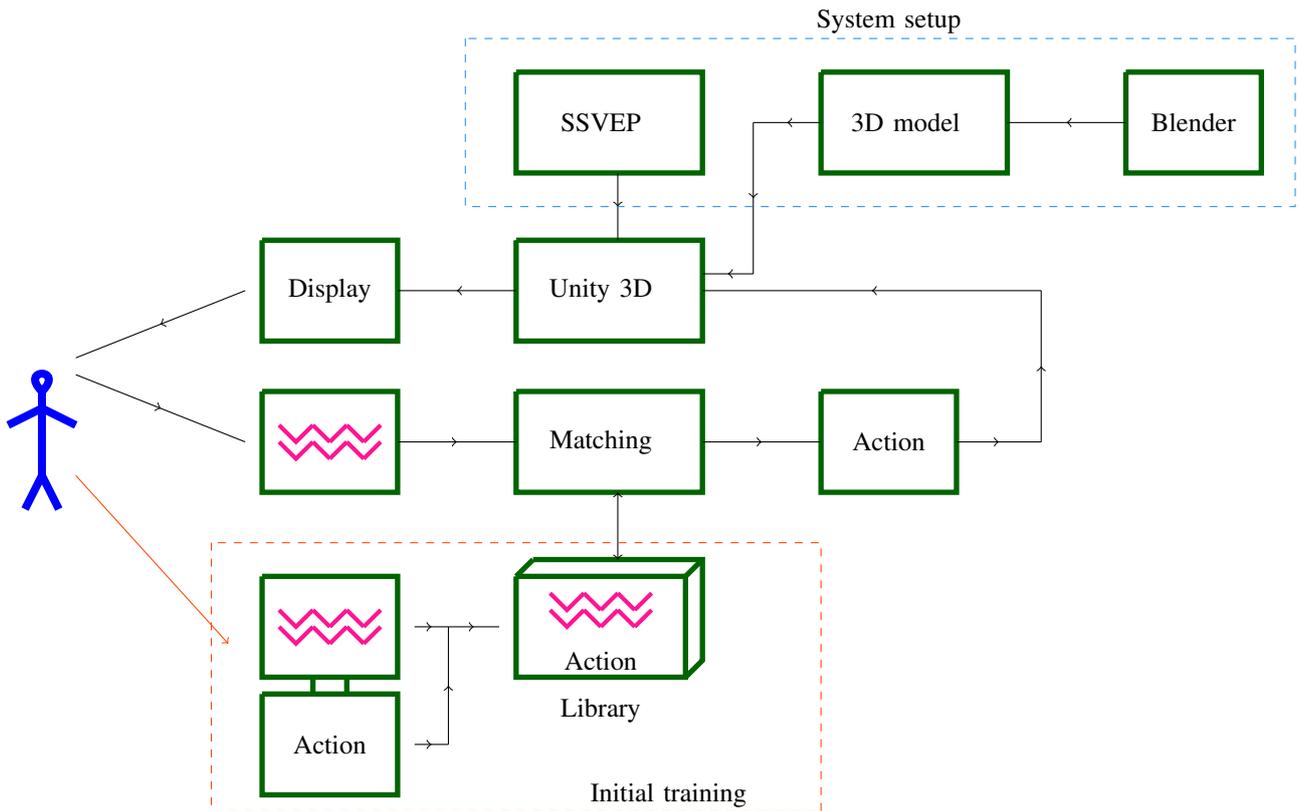


Figure 4: System diagram.

system during development and was able to achieve very precise control of the virtual hand only by using EEG brain waves, that is, without simultaneously moving his physical hands. Had the participant added physical movement we would have likely observed stronger EEG activation, easier brain wave pattern recognition, and even better control. Unfortunately, this was never tested.

To demonstrate proof-of-concept, we report below on the effects of including SSVEP on EEG activation and on game completion time for this single participant. We emphasise, however, that much more testing, under guidance of medical expertise, with a set of different exercises and execution regimes, and with many more participants, both healthy and partially monoplegic patients, is needed before any scientific conclusions can be made.

Effect of SSVEP on EEG Activation

Fig. 5 shows the typical effect on EEG activation with and without SSVEP that we observed for our single participant while performing the stroke rehabilitation game exercise. The effect was present both in early and late stages of testing. The magnitude of brain activity at brain locations is indicated by the spectrum of colours, with the highest activity shown in red and the lowest in blue.

Without the SSVEP, the AF4 and F8 (right frontal lobe) and FC5 (left frontal lobe) regions were the most active in both the alpha and beta bands.

Employing the SSVEP raised the EEG activity markedly across the entire scalp. For alpha waves with SSVEP, the increase in activity was greatest in the right frontal lobe

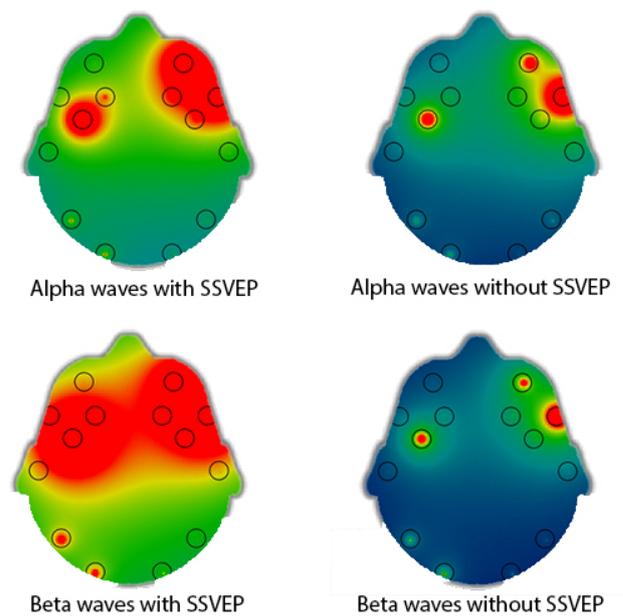


Figure 5: EEG activation with and without SSVEP.

in the AF4, F4, F8, FC6 regions but also the FC5 activity in the left frontal lobe increased. For beta waves with SSVEP, there was a great increase in activity across both frontal lobes (particularly left regions F3, F7, FC5 and

right regions AF4, F4, F8, FC6) but also a marked increase in P7 (left parietal lobe) and O1 (left occipital lobe).

Effect of SSVEP on Game Completion Time

The participant had been using the rehabilitation system extensible before the trials and the effect of improvement due to practice during trials can likely be neglected. The rehabilitation game was configured to finish upon completion of four successful reach-and-grasp exercises. A single reach-and-grasp exercise typically took about half a minute, hence, completing a game would typically take about two minutes. The effect of SSVEP on game completion time is summarised in Table 1.

Trial	Without SSVEP (mm:ss)	With SSVEP (mm:ss)
1	02:04	01:43
2	01:58	01:46
3	02:34	02:06
4	02:16	01:55
5	01:52	01:59
Average	02:09	01:54
St. dev.	00:16	00:09

Table 1: Game completion times with and without SSVEP.

The participant completed the rehabilitation game five times with SSVEP turned on and five times with SSVEP turned off. With SSVEP turned on, the completion time was lower in four trials, whilst the average completion time was 01:54 with a standard deviation of 9 sec. With SSVEP turned off, the average completion time was 02:09 with a standard deviation of 16 sec. Hence, with SSVEP turned on, the average completion time was improved by 15 sec, corresponding to a 12% reduction.

The measurements may suggest that using SSVEP reduces game completion time and its variability as observed by the standard deviation and average completion time, respectively, however, we strongly emphasise that these results are obtained from a single participant and are meant only as an encouragement for further studies when stroke rehabilitation system has been largely improved.

DISCUSSION

This paper has demonstrated fast prototyping of a stroke rehabilitation system consisting of an interactive 3D virtual reality computer game environment interfaced with an EEG headset for control and interaction using brain waves. The system is intended for training and rehabilitation of partially monoplegic stroke patients, was developed over the course of only a few months, and uses low-cost COTS products like the Emotiv EPOC EEG headset and the Unity 3D game engine.

Preliminary testing of a first prototype of the system highlights its potential, however, in order to validate its efficiency in stroke rehabilitation, the system must first be adjusted and improved in close cooperation with medical experts and then run through an extensive field test with a sufficiently large group of patients for comparison with a control group.

COMPARISON WITH EXISTING METHODS

A number of rehabilitation methods exist that can improve motor control and function of the paretic upper limb in partially monoplegic stroke survivors. Unfortunately, most of these methods are commonly characterised by a number of drawbacks that can limit intensive treatment, including being repetitive, uninspiring, and labour intensive; requiring one-on-one manual interaction and assistance from a therapist, often for several weeks; and involve equipment and systems that are complex and expensive and cannot be used at home but only in hospitals and institutions by trained personnel. Mirror therapy, on the other hand, is a simple, inexpensive, and patient-directed rehabilitation method that has been shown to improve hand functioning when used in conjunction with conventional stroke rehabilitation programmes.

Inspired by the principles of mirror therapy, we have developed a game-like computer application in which we deprive the patient of the view of the paretic hand while being challenged with controlling a virtual hand in a 3D game environment only by means of EEG brain waves interfaced with the computer.

We hypothesise that by adopting a rehabilitation scheme similar to mirror therapy, where patients try to physically move their paretic and nonparetic hands whilst using EEG brain waves to control the virtual hand, patients may improve motor control and functioning of their paretic hand.

FUTURE WORK

Whilst our system is only a first prototype, we hypothesise that by iteratively improving its design through refinements and tuning based on input from domain experts and testing on real patients, the system can be tailored for being used together with a conventional rehabilitation programme to improve patients' ability to move the paretic limb much in the same vain as mirror therapy.

Such an improved system would have several advantages over mirror therapy and other conventional rehabilitation methods, namely being (i) game-based and immersive, counteracting laborious and repetitive training exercises and providing a rewarding environment that strengthens and improves rehabilitation; (ii) customisable, allowing for a library of different training exercises not limited by physical equipment; (iii) adaptive and stand-alone, removing the need for instructions from a therapist as the patient progresses and different exercises must be performed; and (iv) extendable, for example by interfacing to a robotic exoskeleton on the paretic limb. In addition, when compared with conventional rehabilitation methods, our system is extremely low-cost and flexible, in particular because patients can use it in the comfort of their homes, with little or no need for professional human assistance.

With respect to extending the system with external physical devices, using Unity as virtual environment enables initial virtual prototyping of the devices. That is, a device first be first modelled, simulated, and interfaced and

controlled by brain waves inside Unity, before an actual physical device is built and connected to the system.

For example, we could provide the user with valuable kinesthetic feedback by connecting the rehabilitation system to a hand exoskeleton or a low-cost haptic glove, such as the one presented by Sanfilippo, Hatledal and Pettersen (2015). With this integration, the patient's hand may be guided with appropriate force feedback for a more engaging learning experience.

Another example is to use EEG brain waves for controlling a prosthetic hand, similar to the mind-controlled, low-cost modular manipulator system presented by Sanfilippo, Zhang and Pettersen (2015). Such a robotic hand may be used for compensating hand function in chronic stroke patients.

Finally, we would like to draw attention to an accompanying paper we submit concurrently, in which we use a similar system as described here, designed to provide tetraplegic patients a training platform for EEG brain control of a virtual electric wheelchair (Hjørungdal, Sanfilippo, Osen, Rutle and Bye, 2016).

REFERENCES

- Alankus, G., Lazar, A., May, M. and Kelleher, C. (2010), Towards customizable games for stroke rehabilitation, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 2113–2122.
- Ang, K. K., Guan, C., Chua, K. S. G., Ang, B. T., Kuah, C. W. K., Wang, C., Phua, K. S., Chin, Z. Y. and Zhang, H. (2011), A large clinical study on the ability of stroke patients to use an EEG-based motor imagery brain-computer interface, *Clinical EEG and Neuroscience* 42(4), 253–258.
- Ang, K. K., Guan, C., Chua, K. S. G., Ang, B. T., Kuah, C., Wang, C., Phua, K. S., Chin, Z. Y. and Zhang, H. (2010), Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback, Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, IEEE, pp. 5549–5552.
- Brunnstrom, S. (1966), Motor testing procedures in hemiplegia: based on sequential recovery stages., *Physical Therapy* 46(4), 357.
- Brunnström, S. (1970), *Movement therapy in hemiplegia: a neurophysiological approach*, Facts and Comparisons.
- Burke, J. W., McNeill, M., Charles, D. K., Morrow, P. J., Crosbie, J. H. and McDonough, S. M. (2009a), Optimising engagement for stroke rehabilitation using serious games, *The Visual Computer* 25(12), 1085–1099.
- Burke, J. W., McNeill, M., Charles, D., Morrow, P. J., Crosbie, J. and McDonough, S. (2009b), Serious games for upper limb rehabilitation following stroke, Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in, IEEE, pp. 103–110.
- Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G. and Dutoit, T. (2013), Performance of the Emotiv EPOC headset for P300-based applications, *Biomedical Engineering Online* 12(1), 56.
- Emotiv (2016a), Emotiv EPOC Brain Computer Interface and Scientific Contextual EEG, <http://www.emotiv.com>. Specifications pamphlet, accessed 29 February 2016.
- Emotiv (2016b), Emotiv website, <http://www.emotiv.com>. Accessed 2 February 2016.
- Fisher, R. S., Harding, G., Erba, G., Barkley, G. L. and Wilkins, A. (2005), Photic-and pattern-induced seizures: A review for the epilepsy foundation of america working group, *Epilepsia* 46(9), 1426–1441.
- Fok, S., Schwartz, R., Wronkiewicz, M., Holmes, C., Zhang, J., Somers, T., Bundy, D. and Leuthardt, E. (2011), An EEG-based brain computer interface for rehabilitation and restoration of hand control following stroke using ipsilateral cortical physiology, Proc. of the IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBC'2011), pp. 6277–6280.
- Hjørungdal, R.-M., Sanfilippo, F., Osen, O. L., Rutle, A. and Bye, R. T. (2016), A game-based learning framework for controlling brain-actuated wheelchairs, Proceedings of the 30th European Conference on Modelling and Simulation (ECMS '16) (submitted for publication), Regensburg, Germany.
- Joo, L. Y., Yin, T. S., Xu, D., Thia, E., Chia, P. F., Kuah, C. W. K. and He, K. K. (2010), A feasibility study using interactive commercial off-the-shelf computer gaming in upper limb rehabilitation in patients after stroke, *Journal of Rehabilitation Medicine* 42(5), 437–441.
- Jordan, K. G. (2004), Emergency EEG and continuous EEG monitoring in acute ischemic stroke, *Journal of Clinical Neurophysiology* 21(5), 341–352.
- Klimesch, W. (1999), EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, *Brain Research Reviews* 29(2), 169–195.
- Kwakkel, G., Wagenaar, R. C., Twisk, J. W., Lankhorst, G. J. and Koetsier, J. C. (1999), Intensity of leg and arm training after primary middle-cerebral-artery stroke: a randomised trial, *The Lancet* 354(9174), 191–196.
- Lewis, G. N., Woods, C., Rosie, J. A. and McPherson, K. M. (2011), Virtual reality games for rehabilitation of people with stroke: perspectives from the users, *Disability and Rehabilitation: Assistive Technology* 6(5), 453–463.
- Masiero, S., Celia, A., Rosati, G. and Armani, M. (2007), Robotic-assisted rehabilitation of the upper limb after acute stroke, *Archives of Physical Medicine and Rehabilitation* 88(2), 142–149.
- Misulis, K. E., Fakhoury, T. and Spehlmann, R. (2001), *Spehlmann's evoked potential primer*, Butterworth/Heinemann.
- Niedermeyer, E. and da Silva, F. L. (2005), *Electroencephalography: basic principles, clinical applications, and related fields*, Lippincott Williams & Wilkins.
- Nudo, R. J. and Duncan, P. W. (2004), Recovery and rehabilitation in stroke introduction, *Stroke* 35(11 suppl 1), 2690–2690.
- Pfurtscheller, G., Flotzinger, D. and Kalcher, J. (1993), Brain-computer interface: a new communication device

- for handicapped persons, *Journal of Microcomputer Applications* 16(3), 293–299.
- Platz, T., Eickhof, C., Van Kaick, S., Engel, U., Pinkowski, C., Kalok, S. and Pause, M. (2005), Impairment-oriented training or bobath therapy for severe arm paresis after stroke: a single-blind, multicentre randomized controlled trial, *Clinical Rehabilitation* 19(7), 714–724.
- Ring, H. and Rosenthal, N. (2005), Controlled study of neuroprosthetic functional electrical stimulation in subacute post-stroke rehabilitation, *Journal of Rehabilitation Medicine* 37(1), 32–36.
- Sanfilippo, F., Hatledal, L. I. and Pettersen, K. Y. (2015), A fully-immersive haptic-audio-visual framework for remote touch, Proc. of the 11th IEEE International Conference on Innovations in Information Technology (IIT'15), Dubai, United Arab Emirates.
- Sanfilippo, F., Zhang, H. and Pettersen, K. Y. (2015), The new architecture of ModGrasp for mind-controlled low-cost sensorised modular hands, Proc. of the 2015 IEEE International Conference on Industrial Technology (ICIT'2015), Seville, Spain, pp. 524–529.
- Stroke, Cerebrovascular Accident* | World Health Organization (2016), http://www.who.int/topics/cerebrovascular_accident/en/. Accessed on 12 February 2016.
- Summers, J. J., Kagerer, F. A., Garry, M. I., Hiraga, C. Y., Loftus, A. and Cauraugh, J. H. (2007), Bilateral and unilateral movement training on upper limb function in chronic stroke patients: a TMS study, *Journal of the Neurological Sciences* 252(1), 76–82.
- Thieme, H., Mehrholz, J., Pohl, M., Behrens, J. and Dohle, C. (2013), Mirror therapy for improving motor function after stroke, *Stroke* 44(1), e1–e2.
- Twitchell, T. E. (1951), The restoration of motor function following hemiplegia in man, *Brain* 74(4), 443–480.
- Ueki, M., Linn, F. and Hossmann, K.-A. (1988), Functional activation of cerebral blood flow and metabolism before and after global ischemia of rat brain, *Journal of Cerebral Blood Flow & Metabolism* 8(4), 486–494.
- Vogiatzaki, E. and Krukowski, A. (2014), Serious games for stroke rehabilitation employing immersive user interfaces in 3D virtual environment, *Journal of Health Informatics* 6, pp. 105–113.
- What is Stroke?* | National Stroke Association (2016), <http://www.stroke.org/understand-stroke/what-stroke>. Accessed on 12 February 2016.
- Yavuzer, G., Selles, R., Sezer, N., Sütbeyaz, S., Bussmann, J. B., Köseoğlu, F., Atay, M. B. and Stam, H. J. (2008), Mirror therapy improves hand function in subacute stroke: a randomized controlled trial, *Archives of Physical Medicine and Rehabilitation* 89(3), 393–398.
- Yavuzer, G., Senel, A., Atay, M. and Stam, H. (2008), "Playstation EyeToy Games" improve upper extremity-related motor functioning in subacute stroke: a randomized controlled clinical trial, *European Journal of Physical and Rehabilitation Medicine* 44(3), 237–244.
- Zhu, D., Bieger, J., Molina, G. G. and Aarts, R. M. (2010), A survey of stimulation methods used in SSVEP-based BCIs, *Computational intelligence and neuroscience* 2010, 1.

AUTHOR BIOGRAPHIES

TOM VERPLAETSE received the BSc degree in electromechanical engineering from Ghent University and is currently studying for a MSc in automation engineering at the same university.

FILIPPO SANFILIPPO³ received the BSc degree in computer engineering from the University of Catania, Catania, Italy, in 2009 and the MSc degree in computer engineering from the University of Siena, Siena, Italy, in 2011. In 2008, he was a Visiting Scholar at the School of Computing and Intelligent Systems, University of Ulster, Londonderry, United Kingdom and in 2010 a Visiting Fellow at the Technical Aspects of Multimodal Systems (TAMS) research group, Department of Mathematics, Informatics and Natural Sciences, University of Hamburg, Hamburg, Germany. In 2015, he received a PhD degree from the Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. For his PhD studies, he was awarded a research scholarship from the IEEE Oceanic Engineering Society (OES) Scholarship program. He is currently working as a Post-Doctoral Researcher at the Department of Engineering Cybernetics, NTNU in Trondheim, Norway. His research interests include control methods, robotics, artificial intelligence and modular robotic grasping.

ADRIAN RUTLE⁴ holds PhD and MSc degrees in Computer Science from the University of Bergen, Norway. Rutle is an associate professor at the Department of Computing, Physics and Mathematics at the Bergen University College, Norway. Rutle's main interest is applying theoretical results from the field of model-driven software engineering to practical domains and has expertise in the development of formal modelling frameworks and domain-specific modelling languages. He also conducts research in the fields of modelling and simulation for virtual prototyping purposes.

OTTAR L. OSEN is MSc in Cybernetics from the Norwegian Institute of Technology in 1991. He is the head of R&D at ICD Software AS and an assistant professor at NTNU in Ålesund.

ROBIN T. BYE⁵ 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 NTNU in Ålesund (formerly Aalesund University College) as a researcher in 2008 and has since 2010 been an associate professor in automation engineering. His research interests belong to the fields of artificial intelligence, cybernetics, and neuroengineering.

³filipposanfilippo.inspitivity.com

⁴www.rutle.no

⁵www.robinbye.com

A Game-based Learning Framework for Controlling Brain-Actuated Wheelchairs

Rolf-Magnus Hjørungdal*, Filippo Sanfilippo[†], Ottar L. Osen*, Adrian Rutle[‡] and Robin T. Bye*

* Software and Intelligent Control Engineering Laboratory
Faculty of Engineering and Natural Sciences

Norwegian University of Science and Technology
NTNU in Ålesund, Postboks 1517, NO-6025 Ålesund, Norway

Email: rolf.hjrdal@gmail.com, {robin.t.bye, ottar.l.osen}@ntnu.no

[†] Department of Engineering Cybernetics

Norwegian University of Science and Technology
NO-7491 Trondheim, Norway

Email: filippo.sanfilippo@ntnu.no

[‡] Department of Computing, Mathematics and Physics

Bergen University College

P.O. Box 7030, Bergen, Norway

Email: adrian.rutle@hib.no

KEYWORDS

Brain-computer interface, electroencephalography, virtual reality, low-cost commercial-off-the-shelf products.

ABSTRACT

Paraplegia is a disability caused by impairment in motor or sensory functions of the lower limbs. Most paraplegic subjects use mechanical wheelchairs for their movement, however, patients with reduced upper limb functionality may benefit from the use of motorised, electric wheelchairs. Depending on the patient, learning how to control these wheelchairs can be hard (if at all possible), time-consuming, demotivating, and to some extent dangerous. This paper proposes a game-based learning framework for training these patients in a safe, virtual environment. Specifically, the framework utilises the Emotiv EPOC EEG headset to enable brain wave control of a virtual electric wheelchair in a realistic virtual world game environment created with the Unity 3D game engine.

INTRODUCTION

The ability to move around, explore our surroundings and being able to transfer to other places in order to take part in daily activities is an essential quality in human life. People with disabilities may lack this ability due to their illness. With the aid of prostheses or wheelchairs, many disabled people can become more mobile. However, it may be very difficult or even impossible for tetraplegic patients, or paraplegic patients with reduced upper limb functionality, to control an electric wheelchair via a joystick or other manual control devices. For these patients, an electric wheelchair that can be operated solely by the mind

could provide a formidable improvement in the quality of life.

Prior to the development of brain-actuated wheelchair, several factors must first be considered. In order to map out the needs and shortcomings of the available technology it is desirable to test the existing technology in a virtual environment. This enables the exploration, testing and development of the user interface and brain-computer interface (BCI) functionality.

This paper presents the development of an open-source framework for disabled people such as tetraplegics or sufferers of amyotrophic lateral sclerosis (ALS) to control a brain-actuated wheelchair in a virtual environment. The framework is realised by exclusively adopting low-cost commercial off-the-shelf (COTS) components and tools. In particular, an electroencephalography (EEG) headset, the Emotiv EPOC, is chosen for monitoring the subject's brain waves. These signals are then used as inputs for controlling a wheelchair in a simulated environment. To achieve this goal, the Unity 3D game engine is selected as an efficient integration platform. The adopted design choices make the proposed framework very flexible and extremely low-cost.

In the following sections, we first provide a background of game-based learning, EEG technology, related work, and our motivation and aim. We proceed with describing our game-based methodology before presenting the framework architecture, including the Emotiv headset and library and the Unity environment, the interface between the headset and Unity, design of the game environment, and a preliminary artificial neural network (ANN) (Yegnanarayana, 2009) for converting raw EEG brain waves into control signals. Finally, we present the results of our work and a discussion, including future work.

Corresponding author: Robin T. Bye.

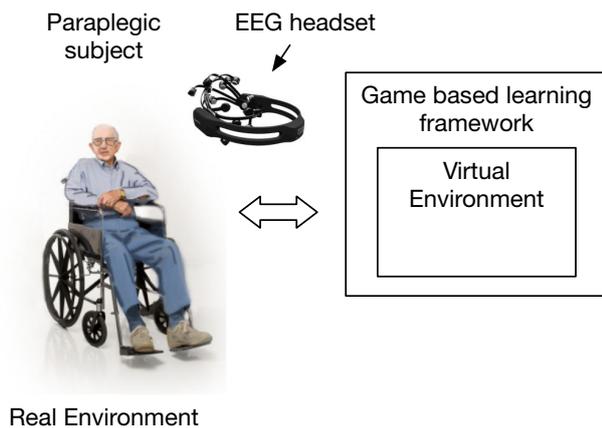


Figure 1: Game-based learning framework for paraplegic subjects to control brain-actuated wheelchairs.

BACKGROUND

There is a rapidly growing body of empirical evidence on the effectiveness of using video and computer games to provide instruction (Tobias et al., 2014). We experience pleasure when actively engaged in games, especially in coming to understand how a new system works. This is true whether the game is considered “entertainment” or “serious” (Susi et al., 2007). Here, we adopt a game-based learning approach, where a “serious” game is developed to engage disabled subjects in learning how to control their wheelchair by using brain waves. We believe our approach is especially suitable for this kind of training, because mastering brain wave control can be difficult and involve tedious and repetitive tasks and thus be demotivating for the subject. Figure 1 illustrates the main idea and a high level view of the components involved in the game-based learning framework.

EEG Technology

EEG is an electrophysiological monitoring method that measures the natural electric potential on the scalp (Niedermeyer and da Silva, 2005). Physiologically, EEG power reflects the number of neurons that discharge synchronously (Klimesch, 1999). This electric potential is a result of brain activity and behaves in a periodic, wavelike fashion referred to as brain waves and can be recorded with a portable EEG headset such as the Emotiv EPOC EEG (e.g., Duvinage et al., 2013).

Traditionally, EEG technology has been a diagnostic tool for medical professionals in order to diagnose neurological disorders such as epilepsy, brain tumors and more (Vaque, 1999). In recent years, EEG technology has seen a commercialisation which has resulted in affordable EEG equipment for both researchers and end-consumers. One such device is the Emotiv EPOC headset, which comes with several pieces of software developed for BCI. The included software enables the user to both record and investigate their brain wave activity in real time with the brainwave signals split into the conventional EEG frequency bands, namely the delta (< 4 Hz), theta (4–7

Hz), alpha (8–15 Hz), beta (16–31 Hz), and gamma (> 32 Hz) bands, of which the alpha band, which is active during an alert and cognitive state of the patient (Klimesch, 1999), and the beta band, which is closely related to purposive movement (Niedermeyer and da Silva, 2005), are the most important with respect to BCI.

One of the features of this software is the ability to store specific brain wave patterns as commands. Up to four different user-defined commands can be stored (e.g., up, down, left, right), and the proper software for mapping and training these commands are provided. The application programming interface (API) provided by Emotiv enables software developers to access these commands and use them in their software, achieving BCI as a result.

Related Work

The possibility of using EEG technology for medical assistant applications has been studied in the literature. In particular, the possibility of enabling disabled subjects to control their wheelchairs by using brain waves has been investigated by several research groups. For instance, an attempt to use brain signals to control mechanical devices such as wheelchairs was presented by Tanaka et al. (2005). To achieve this goal, a recursive training algorithm to generate recognition patterns from EEG signals was developed. Relevant experimental results demonstrated the potential of the proposed system. In particular, the system was tested in a real experimental scenario where subjects were required to approach target positions by repeating movements. However, this experimental setup required external assistance for the subjects because the wheelchair was supposed to be stopped during EEG detection and pattern matching, since the processing time was very slow. Successively, thanks to different advances in this technology, a real-time EEG classification system was presented by Craig and Nguyen (2007), with the goal of enhancing the control of a head-movement controlled power wheelchair for patients with chronic spinal cord injury (SCI).

One of the main challenges that characterises most of these previous works is that developing and testing brain-actuated wheelchairs in a real-world environment is very difficult because of the extensive training that is required for the paraplegic subjects to safely operate the systems. In this perspective, even though numerous research efforts have been performed to develop brain-actuated wheelchairs, to the best of our knowledge there exists only a few integrated frameworks for effectively training paraplegic or tetraplegic subjects in controlling their wheelchairs. For instance, the possibility of a tetraplegic for using brain waves to control movements of his wheelchair in virtual reality was first studied by (Leeb et al., 2007). In this study, a tetraplegic subject was able to generate bursts of beta oscillations in the EEG by imagination of movements of his paralyzed feet. These beta oscillations were used for self-paced (asynchronous) BCI control based on a single bipolar EEG recording. The subject was placed inside a virtual street populated with avatars. Even though the use

of a visual-rich virtual environment was proved to be a very effective approach for improving efficiency of virtual training, the possibility of adding some elements of game-based learning was not deeply considered.

Motivation and Aim

As described above, development and testing of brain-actuated wheelchairs for paraplegic patients in a real-world environment is challenging due to safety concerns. Only recently did researchers attempt to use virtual reality training to overcome this challenge but failed to include the advantage of game-based learning aspects. Moreover, using a virtual environment enables an adaptive and incremental learning paradigm, where training exercises are matched with the level of skill attained by the users. Finally, virtual environments are easy to modify and extend in software compared to their physical counterparts.

Motivated by these factors in an emerging field of research, our aim is to propose a game-based learning framework for brain-actuated wheelchairs, developed in only a few months by a small group of people, and involving only low-cost COTS components, that incorporates many of the advantages a virtual training environment can offer.

METHOD

This section describes the implementation details of our framework, and the rationale for some of our design decisions.

Game-based Methodology

The aim of implementing game-based learning concepts is to create a training environment in such a manner that the trainee is learning while at the same time enjoying the game aspects of the exercises. The skills and knowledge provided by the experience of playing the game can then later be applied in real-life scenarios.

Several aspects of game-based learning have been implemented in our virtual training environment in order to enhance learning, including

- safe risks, where users can experience consequences from their mistakes in a safe environment;
- goal-based tasks, where an on-screen prompt informs the user about the current task;
- incremental learning, where the user is prompted to complete more challenging tasks as the game progresses; and
- timed events, where users can compare themselves with their previous times.

Learning to consistently switch between input commands is imperative to safely operate a brain-actuated wheelchair. In our virtual environment, the user can explore and test the wheelchair functionality in a setting free from risk in order to improve their BCI skills. As a result, a patient utilizing a brain-actuated wheelchair in the future may greatly benefit from the experience gained in the game-based learning environment.

Users can also benefit from competing against themselves, for example trying to reduce their completion time for a particular game level. A shorter completion time would likely be due to better ability to switch input commands.

We have attempted to address some of the advantages of a game-based methodology in our implementation. For example, we have included the abovementioned game completion timer that accompanies an onscreen description of the task at hand where it is relevant. Furthermore, we have created an environment with an emphasis on step-by-step self-paced incremental learning, where users complete game levels with progressively more difficult tasks, exploring different aspects of the skills needed to operate an electric wheelchair in a real-world environment.

The proposed game-based training methodology follows the game levels depicted in Figure 2. In Level 1, users begin by concentrating on only learning and practicing a single input command, namely that of moving forward in a “drag race,” in which the task is to drive the wheelchair straight forward from the starting line to the finish line. After this level has been completed, the player can choose to repeat the level in order to improve completion time, or to proceed to more challenging tasks involving several input commands. Specifically, in Level 2, the user will learn to switch between movement commands in order to navigate a labyrinth. In Level 3, users learn to handle safety mechanisms, whereas Level 4 offers more difficult and advanced tasks for improving navigation skills. Finally, Level 5 provides users with a realistic real-world scenario, where users must navigate the wheelchair in an urban area, using all the skills they have learnt previously.

Framework Architecture

The framework consists of two main components: the EEG headset and the game engine. The choice of these two components is crucial for the success of the framework. In this section, each of the main components are explained and finally the interfacing between the game engine and the EEG device is described in detail.

Emotiv Headset and Library: The EEG device connects the user’s brain to the virtual environment by converting EEG brain waves measured from one or several electrodes positioned on the user’s scalp into BCI commands. There were two factors that we considered the most important when choosing the EEG hardware to achieve our goals, namely convenience and accuracy.

In terms of convenience, the device must be simple for the end user to equip and use. In addition the device should have a good API in order to make it easy for the developer to create software.

In terms of accuracy, it is important that the EEG device provides enough accuracy to differentiate between several mental states for control, e.g. “forward,” “left,” and “right.” This is necessary in order for the user to be able to move around freely in an open, unconstrained environment such as the real world.

Compared to most conventional medical equipment, the Emotiv EPOC is very convenient to use as it is intended

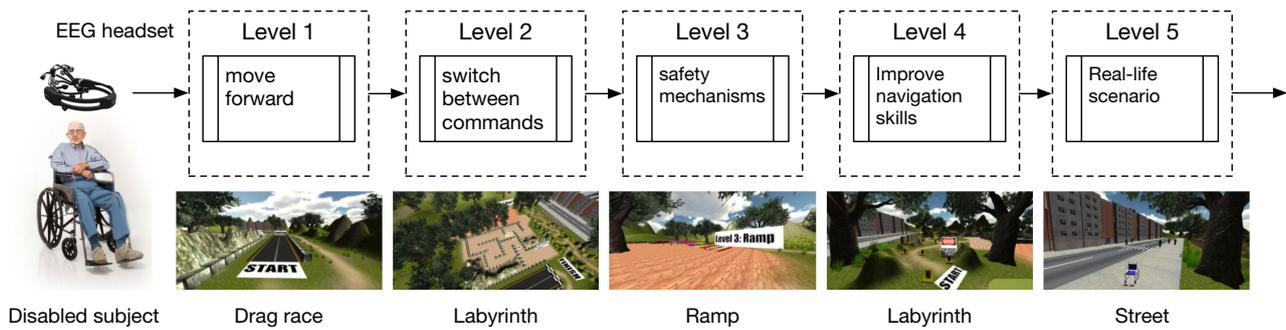


Figure 2: The proposed game-based training methodology. A training sequence of different levels is adopted.

for end-consumers and does not require professional expertise. Considering the nature of how EEG is measured, convenience is not compatible with accuracy and some sacrifices must be made in order to ensure the quality of the measured signals. Nevertheless, compared to other COTS EEG equipment that solely uses dry electrodes, the Emotiv EPOC is designed to be used with a saline solution that is applied to electrodes, thus significantly improving the connection between the capacitive sensors and the scalp compared to dry electrodes.

Furthermore, the numerous electrodes of the Emotiv EPOC along with its advanced software for brain wave pattern recognition enables the device to recognise up to four different mental states that can be used simultaneously for EEG brain control. Whilst the source code of Emotiv software is proprietary and hidden, it is clear that the recognition of these mental states is presumably done by means of some advanced machine learning algorithm such as an ANN and is conveniently available to programmers via the API.

Also convenient is the availability of an official plugin compatible with Unity 4.6 that connects the Emotiv API to Unity. This enables software developers to implement in Unity the features provided by the headset with ease. Hence, considering these facts, the Emotiv EPOC provides a solid middle ground between convenience and accuracy.

Unity Environment: A number of 3D game engines for developing virtual environments exist, with perhaps some of the most popular being Unity, Unreal Engine, CryEngine, and Source 2. While each of these have their own strengths and weaknesses, we wish to highlight that the engine of our choice, Unity, is very quick to learn whilst simultaneously being quite advanced, thus enabling very fast prototyping of virtual game environments. Another strength of the Unity game engine is the cross-platform focus of the engine, which potentially enables portability between platforms such as tablets or smart phones in future development. Finally, as stated above, Emotiv has developed an official Unity plugin which further simplifies the relationship between the headset and the game engine. By having access to a readily available plugin to the Emotiv API, choosing the Unity engine ensures that a minimal amount of time is used to interface the EEG device to the game engine and more time can be used to develop the virtual environment.

An illustration of the framework architecture is shown in Figure 3. The Emotiv “EmoEngine” handles the EEG signals from the headset and interfaces with the Emotiv API, enabling programmers to access both raw and processed EEG signal, thus utilizing BCI functionality. The Unity plugin provides an interface between the Emotiv API and the Unity game engine, which is used for creating the virtual training environment. More details on the components and their interaction are provided in the following sections.

Interfacing Emotiv EPOC with Unity

Before the EPOC headset can be used as a BCI, the user must record at least one mental state associated with a “BCI command.” This is done via the bundled Emotiv Control Panel. This software stores the user profiles and their relationship between mental states and BCI commands. When the configuration process begins, the user is prompted for a “neutral” state of mind for a short amount of time in order for the algorithm to have a neutral base where no command is active. After the neutral state is stored, the user may proceed to configure up to four other commands, in our case labelled by the Emotiv software as “push,” “pull,” “left,” and “right.” These four commands can later be accessed programmatically from within Unity via the plugin. After the desired BCI commands have been recorded, the user can attempt to activate them again inside the wheelchair framework. Importantly, these commands can be mapped to have different meanings in different setting, e.g., “push” can be used as a “forward” command in one particular setting, or mode, while being used as a “choose” command in another mode.

When the headset measures an EEG state similar to a previously recorded EEG state, the EPOC determines how closely it matches the original recording, and assigns it a number in the interval $[0, 1]$, where 1 means a perfect match. Algorithm 1 shows how this information can be accessed in Unity. This particular piece of code is executed at a rate of 60 times per second, and queries the Emotiv plugin for updated information.

When recording the BCI commands it may be beneficial for the user to mentally associate the commands with physical movements. For example, one may associate “push” with walking, and “left” and “right” with movement of the left and right arms, respectively, thus making it easier to reproduce a particular BCI command. However, in cases

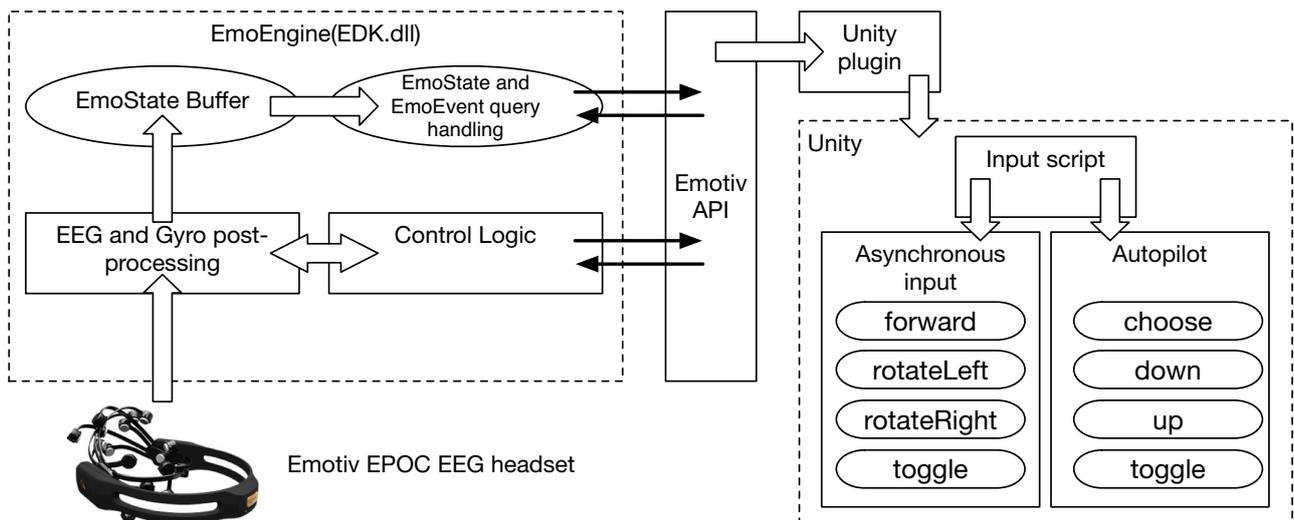


Figure 3: The proposed framework architecture. Some elements of this figure are credit to Emotiv (Emotiv, 2014).

```

void CognitivActionUpdate() {
    int count=0;
    foreach (float f in EmoCognitiv.
        CognitivActionPower) {
        //only the current active command can be
        //greater than zero
        if (f>0f) {
            CurrentCognitivPower = f;
            CurrentCognitivAction =
                EmoCognitiv.cognitivActionList[count].
                ToString();
        }
        else count+=1;
    }
}

```

Algorithm 1: Method of obtaining cognitive data from the plugin.

where the person never had the ability to walk or make arm movements, slightly more creative approaches must be made. For example, some testing was done using various mental images unrelated to bodily movement. An example of this could be to imagine a cube suspended by rubber bands inside one’s own head. To imagine the movement or rotation of such a cube will require mental concentration, which in turn affects the EEG state of the user that can be used as a control signal.

Design of Game Environment

This section describes how the environment was created together with the design choices underneath the surface. The aim was to create an open game world where the user can roam freely around the environment whilst having terrain, trees, rubble and buildings limit movement to a reasonable degree. This is done in order to both create an intuitive understanding of what to do and where to go, and at the same time avoid the artificial feeling one might get from a minimalistic design using invisible borders.

The world is divided into five game levels as described previously (see Figure 2), each testing various

BCI commands and implemented functions as illustrated in Figure 3.

The wheelchair has two modes of operation that the user can switch between at any time, a manual self-paced asynchronous mode (as opposed to cue-based synchronous mode) and an autopilot mode. We mainly adopt the term “asynchronous” in this paper since this term is commonly used in the literature but for most purposes the term is equivalent to “manual,” meaning that the user is not limited by any cues or overridden by an autopilot, but is free to make movements at will.

In the asynchronous mode, the user can move around freely while learning and practicing BCI commands by completing tasks prompted by a context-sensitive graphical user interface (GUI) (an example is shown in Figure 4).

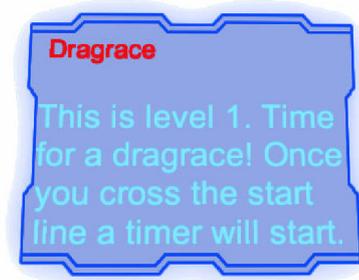


Figure 4: Context-sensitive GUI describing the current objective.

The autopilot mode enables the user to travel between five predefined geographical locations in the virtual environment. The autopilot is realized by utilizing the A*-algorithm (Hart et al., 1968) for pathfinding, and is accessed via the GUI. For real-world purposes, the autopilot would have to incorporate real maps and a means to select target locations, for example by freely available map services online. Great care must also be taken to ensure that the chosen path is safe and accessible for an electric

wheelchair.

Being limited to four degrees of freedom impose some challenges when designing how the user should interact with the application, when there are two modes of operation involved. Several designs were considered prior to choosing the design illustrated in Figure 3. The diagram illustrates the inner workings of the Emotiv software engine, and how the virtual environment is connected. The “EmoEngine” receives pre-processed EEG and gyro data from the headset which will then be post-processed into an “EmoState”-structure that contains the currently active BCI command. These structures can be accessed via an “EmoEvent” query which will return the current state. In this case, the querying for an “EmoEvent” is handled by the Emotiv API and the Unity plugin (Emotiv, 2014).

The number of mental commands that the Emotiv software can learn and store for a particular user is limited to four. However, by using modes, we can use these four commands for many different purposes. Specifically, the function activated by a BCI command depends on which mode is currently active, and whether the command’s threshold value has been exceeded.

In asynchronous mode, *forward*, *rotateLeft*, *rotateRight* and *toggle* are activated by the BCI commands “push,” “left,” “right,” and “pull,” respectively. The purpose of the first three commands is to move the wheelchair forward or rotate it to the left or right, whilst the last command is used to switch to autopilot mode and back.

In autopilot mode, the wheelchair is operated in a similar fashion, with the BCI command “push” mapped to *choose*, “left” is mapped to *down*, “right” is mapped to *up*, and “pull” is again mapped to *toggle*. The user uses the *up* and *down* commands to navigate a list of destinations, and then selects it using the *choose* command.

Importantly, when having two or more modes in the game, every mode must include a command for changing modes. Here, we use a *toggle* but for more complex structures, possibly involving many modes and even submodes, a better command could be *back*, which is well known to users of smartphones and tablets.

Preliminary Artificial Neural Network (ANN)

Whilst the Emotiv software for brain wave pattern recognition is a powerful tool for the framework we describe here, it is proprietary and closed source. This may limit the functionality of the framework and also forces it to be compatible with current and future versions of the Emotiv software. We therefore decided to implement a preliminary ANN for EEG pattern recognition and classification. In machine learning and cognitive science, ANNs are a family of models inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown (Yegnanarayana, 2009). This technology is particularly relevant for the application recognising mental BCI commands, with a large number of inputs that need to be considered when acquiring data with an EEG headset.

Here, an experiment was performed to investigate whether an ANN was able to classify two types of EEG

states, namely “meditation” and the command “push.” We collected 200 raw EEG data samples comprised of these two different cognitive states. The first 100 samples were sampled while the user was meditating with closed eyes, trying to become completely relaxed. The other 100 samples were sampled while the user was trying to generate the cognitive action “push,” which can be considered a polar opposite to “meditation.” We used the Neural Network Toolbox in Matlab (Mathworks, Inc., 2015) to implement the ANN and perform the experiment.

During the EEG data collection for the experiment, the user controlled the wheelchair in Unity to ensure that the correct cognitive state was activated, meaning that if the wheelchair was not moving, the “push” sample would be discarded. Measuring the quality of meditation is usually not a straightforward process because the nature of high quality meditation remains subjective. As long as the eyes are closed, less beta activity in the brain can be normally expected. For these reasons and in order to ensure good reliability and accuracy for the considered data set, each data sample was sampled with a duration of 10 seconds by the same user during the same day.

As inputs to the ANN, the mean power spectral density from seven EEG channels was used. The power spectrum was further divided into six frequency bands, namely the delta, theta, and alpha bands, and the low, medium, and high subbands of the beta band. Using some rules-of-thumb and trial-and-error, the number of hidden neurons was set to 21 for best results. Summarising, the ANN therefore consisted of a 6×7 inputs, 21 hidden neurons, and an output categorising the input as either “meditation” or “push”, as depicted in Figure 5. The sample size of the training set was 140, whereas both the validation set and testing set was set to 30 samples.

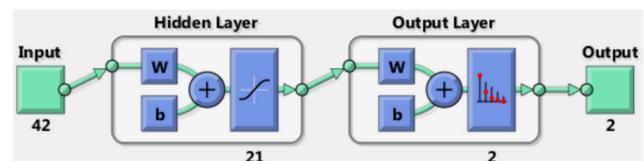


Figure 5: Structure of artificial neural network (ANN) for classification of two EEG states.

RESULTS

A large and realistic urban 3D virtual world has been implemented, in which users can control a virtual brain-actuated wheelchair to navigate this world. Within this virtual environment, there are five incrementally more difficult game levels for game-based learning and practicing of brain control of the wheelchair (see Figures 6–10 for screenshots).

Three young and healthy male students in their twenties using the virtual training environment were all able to learn how to control the virtual wheelchair only with their minds. In particular, the students were able to utilise all the different BCI commands to complete all game levels successfully with good scores in a self-paced asynchronous



Figure 6: Level 1.



Figure 7: Level 2.



Figure 8: Level 3.



Figure 9: Level 4.



Figure 10: Level 5.

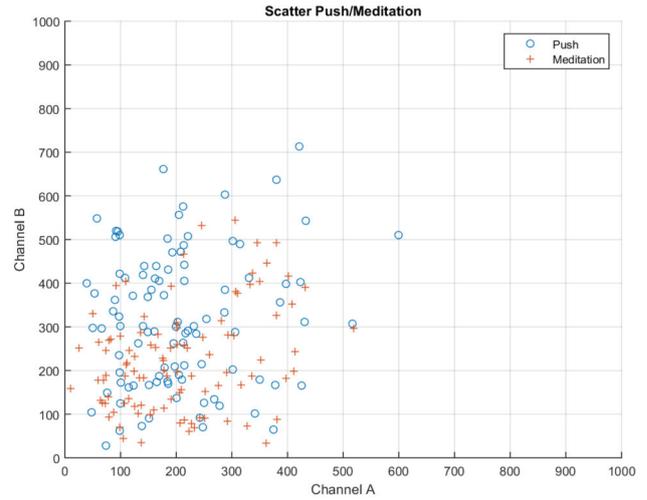


Figure 11: Scatter plot of EEG amplitudes for channel 4 (A) and 7 (B) at 20Hz.

mode; to toggle between asynchronous mode and autopilot mode; and to tell the autopilot to plan and move the wheelchair to one of five geographical locations in the virtual world, all only by means of EEG brain waves.

Lower completion time of a game level is equivalent to better control of the brain-actuated wheelchair. A timer system was used that both displays the amount of time the user have spent in a GUI as well as stores the completion time in a high score system.

Two safety features were implemented, namely collision avoidance and rolling protection. Navigating the wheelchair, if suddenly on collision course with an object tagged as “collidable” (which is almost any object in the virtual environment), the safety brakes will engage if the wheelchair gets too close. In addition, brakes are also engaged to avoid unwanted rolling if the wheelchair’s angle relative to the horizontal plane reaches a certain threshold and no control command is present.

The A* algorithm was used as basis both for path planning in the autopilot mode, and for simulating people walking around on streets.

Preliminary ANN experiment

To gain some insight into the acquired EEG data, a scatter plot of the signal amplitude of EEG channels 4 and 7 at 20 Hz for either “meditation” samples or “push” samples is depicted in Figure 11. The two cognitive states are not aggregated in clusters, which means that observing EEG at just one particular frequency generally is not sufficient for classification of a BCI command. It is for this reason we had to use bands of frequency with mean power spectral density instead. As a matter of fact, running the ANN with a set of discrete frequencies proved to be insufficient for classifying the two EEG states. Instead, when running the ANN as described previously using the mean power spectral densities for six bands of frequencies, the ANN was able to consistently discern meditation from the “push” command with an error rate of 0–4%,

depending on the initialisation values of the network, and the distribution of samples in the training, validation, and test sets.

DISCUSSION

In this paper, we present a game-based learning framework for controlling brain-actuated wheelchairs. The framework may be used to train paraplegic patients with paralysis of the upper body in a safe virtual environment before introducing them to real wheelchairs. The virtual wheelchair is controlled only using brain waves through the low-cost COTS product Emotiv EPOC headset. Moreover, the virtual environment has been developed using the Unity 3D game engine. Compared to existing works, we use a game-based learning methodology to motivate, enhance and speed up the learning process. The framework is modular and flexible, with easy extensions to more features such as game levels and skill training, multiplayer options, interfacing to a real wheelchair, and more.

Safety and Control

Safety is an extremely important issue before actual real-world brain-actuated wheelchairs can be used. Here, we have implemented two safety features, namely collision avoidance and rolling protection. Collision avoidance would be much harder to do in a real-world, uncertain, dynamic environment than in our controlled virtual world, and would likely require the use of advanced artificial intelligence (AI) for computer vision, planning, and decision-making. Using the rolling protection mechanism for a real wheelchair would need a gyroscope but should be fairly easy to implement for the physical wheelchair.

A safety feature that probably should be implemented include some kind of AI emergency interruption, where the AI overrides the BCI command provided by the user if the execution of the command can be dangerous, such as driving the wheelchair into a street with heavy traffic.

It may also be that the inclusion of a third semi-autonomous, or cue-paced (synchronous), mode could improve control and thereby safety. In such a mode, the user is presented with cues such as a visual image, text, sound, or similar for triggering particular BCI commands. An AI decision-support system could infer what would be a good command at a given moment, for example, turning left at an intersection, and present the user with the cue that corresponds to turning left.

Finally, one could take advantage of the steady state visually evoked potential (SSVEP), much like we do in our accompanying paper submitted concurrently (Verplaetse et al., 2016). When providing a user with a computer screen rapidly switching between two colours at a given frequency, one can evoke a SSVEP for higher EEG activation and better mental control. According to a survey by Zhu et al. (2010), BCI systems based on the SSVEP provide a higher level of information throughput and require shorter training than BCI systems using that are not augmented with SSVEP. The SSVEP could be used

in a manner similar as the cue-paced mode above to aid in generating certain EEG patterns and BCI commands at discrete points in time.

Reverse-Engineering Emotiv Software

The preliminary ANN experiment is equivalent to a first baby step towards reverse-engineering the proprietary software developed by Emotiv for generating BCI commands based on EEG signals. We were able to successfully implement a simple ANN able to classify two EEG states: meditation and “push.” The experiment has provided some insight into how we can use ANNs for such brain wave pattern recognition. However, we acknowledge that there is still much work to do, and that the task we adopted was simple. If we had chosen two BCI commands that both require concentration and mental focus on a command (as opposed to meditation, which aims to reduce concentration), it may have been more difficult for the ANN to perform classification. Likewise, with more BCI commands to classify, the problem also becomes harder. Nevertheless, our experiment does seem to indicate that ANNs are suitable for solving this problem.

Future Work

As future work, it would be interesting to consider the possibility of adding some level of adaptability to the proposed game-based methodology to improve the learning experience. This could be achieved by developing a specific learning algorithm that can adapt the level of external assistance provided to the subject according to the subject’s experience. A similar algorithm has been presented by several researchers (Philips et al., 2007; Millan et al., 2009). The underlying idea was to provide the subject with an adaptive level of support, thereby complementing the user’s capabilities at any moment, whenever necessary. An inexperienced user will receive more assistance than an experienced one. If, after some time, the performance of the user has improved, the assisting behaviours will be less activated. By introducing this adaptability, the users remain in maximal control.

To make the game-based learning experience more immersive and therefore even more engaging for the user, the integration with an open-source low-cost framework for a fully-immersive haptic, audio and visual experience like the one proposed by Sanfilippo, Hatledal and Pettersen (2015) may be considered. This framework allows for establishing a kinesthetic link between a human operator interacting with a computer-generated environment.

One more possible future work that we are considering is the possibility of implementing a shared control system between the a simulated and a real wheelchair. The system can then serve the purpose as a platform for virtual prototyping of the real wheelchair, where modelling, features, functionality and so forth can be simulated before the real physical wheelchair is built. This approach may also be very useful for minimising the difficulties for the subjects to switch from a simulated system to a real system when the training programme is terminated. In addition,

comparative studies can be performed concerning usability and taking into account human factors.

The concept of EEG and BCI can probably be beneficial for other human assistance technologies. One exciting application could be that of intelligent prostheses or exoskeletons that likely would require the use of machine learning algorithms and evolutionary computation, with which we have extensive experience at NTNU in Ålesund (e.g., see Sanfilippo et al., 2013, 2014; Sanfilippo, Hatledal, Styve, Zhang and Pettersen, 2015; Bye et al., 2015; Bye and Schaathun, 2015; Alaliyat et al., 2014; Hatledal et al., 2014, for work relating to genetic algorithms, particle swarm optimisation, ANNs, and more).

Other possible work to be considered in the future may include testing of different machine learning algorithms and compare their corresponding performances. In order to do this, a machine learning framework that provides a selection of existing learning approaches and allows for implementing new algorithms can be used as presented in (Hatledal et al., 2014). This framework can be used to develop a standard benchmark suite for testing and measuring the effectiveness and accuracy of the compared methods.

Finally, we would like to draw attention to an accompanying paper we submit concurrently, in which we use a similar system as described here, designed to provide partially monoplegic stroke patients with a rehabilitation platform using EEG brain control of a virtual paretic hand (Verplaetse et al., 2016).

REFERENCES

- Alaliyat, S., Yndestad, H. and Sanfilippo, F. (2014), Optimisation of Boids Swarm Model Based on Genetic Algorithm and Particle Swarm Optimisation Algorithm (Comparative Study), Proceedings of the 28th European Conference on Modelling and Simulation.
- Bye, R. T., Osen, O. L. and Pedersen, B. S. (2015), A computer-automated design tool for intelligent virtual prototyping of offshore cranes, Proceedings of the 29th European Conference on Modelling and Simulation (ECMS'15), pp. 147–156.
- Bye, R. T. and Schaathun, H. G. (2015), A simulation study of evaluation heuristics for tug fleet optimisation algorithms, Operations Research and Enterprise Systems. In Communications in Computer and Information Science, Springer, pp. 165–190.
- Craig, D. A. and Nguyen, H. (2007), Adaptive EEG thought pattern classifier for advanced wheelchair control, Proceedings of the 29th IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBS), pp. 2544–2547.
- Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G. and Dutoit, T. (2013), Performance of the Emotiv Epoc headset for P300-based applications, *Biomedical Engineering Online* 12(1), 56.
- Emotiv (2014), *Emotiv Software Development Kit*. <http://emotiv.com/developer/SDK/UserManual.pdf>
- Hart, P. E., Nilsson, N. J. and Raphael, B. (1968), A formal basis for the heuristic determination of minimum cost paths, *IEEE Transactions on Systems Science and Cybernetics* 4(2), 100–107.
- Hatledal, L. I., Sanfilippo, F. and Zhang, H. (2014), JIOP: a java intelligent optimisation and machine learning framework, Proceedings of the 28th European Conference on Modelling and Simulation (ECMS), Brescia, Italy, pp. 101–107.
- Klimesch, W. (1999), EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, *Brain Research Reviews* 29(2), 169–195.
- Leeb, R., Friedman, D., Müller-Putz, G. R., Scherer, R., Slater, M. and Pfurtscheller, G. (2007), Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic, *Computational Intelligence and Neuroscience* 2007.
- Mathworks, Inc. (2015), *MATLAB Neural Network Toolbox*, The Mathworks, Inc., Natick, Massachusetts.
- Millan, J. D. R., Galán, F., Vanhooydonck, D., Lew, E., Philips, J. and Nuttin, M. (2009), Asynchronous non-invasive brain-actuated control of an intelligent wheelchair, Proceedings of the IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC), pp. 3361–3364.
- Niedermeyer, E. and da Silva, F. L. (2005), *Electroencephalography: basic principles, clinical applications, and related fields*, Lippincott Williams & Wilkins.
- Philips, J., Millán, J. d. R., Vanacker, G., Lew, E., Galán, F., Ferrez, P. W., Brussel, H. V. and Nuttin, M. (2007), Adaptive shared control of a brain-actuated simulated wheelchair, Proceedings of the IEEE 10th International Conference on Rehabilitation Robotics (ICORR), pp. 408–414.
- Sanfilippo, F., Hatledal, L. I. and Pettersen, K. Y. (2015), A fully-immersive haptic-audio-visual framework for remote touch, Proceedings of the 11th IEEE International Conference on Innovations in Information Technology (IIT'15), Dubai, United Arab Emirates.
- Sanfilippo, F., Hatledal, L. I., Schaathun, H. G., Pettersen, K. Y. and Zhang, H. (2013), A universal control architecture for maritime cranes and robots using genetic algorithms as a possible mapping approach, Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO), Shenzhen, China, pp. 322–327.
- Sanfilippo, F., Hatledal, L. I., Styve, A., Zhang, H. and Pettersen, K. Y. (2015), Integrated flexible maritime crane architecture for the offshore simulation centre AS (OSC): A flexible framework for alternative maritime crane control algorithms, *IEEE Journal of Oceanic Engineering* PP(99), 1–12.
- Sanfilippo, F., Hatledal, L. I., Zhang, H. and Pettersen, K. Y. (2014), A mapping approach for controlling different maritime cranes and robots using ANN, Proceedings of the 2014 IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, pp. 594–599.
- Susi, T., Johannesson, M. and Backlund, P. (2007), Serious games: An overview.
- Tanaka, K., Matsunaga, K. and Wang, H. O. (2005), Electroencephalogram-based control of an electric wheelchair, *IEEE Transactions on Robotics* 21(4), 762–766.
- Tobias, S., Fletcher, J. D. and Wind, A. P. (2014), Game-based learning, Handbook of research on educational communications and technology, Springer, pp. 485–503.
- Vaque, T. (1999), The history of EEG Hans Berger: Psychophysicist. A Historical Vignette., *Journal of Neurotherapy: Investigations in Neuromodulation, Neurofeedback and Applied Neuroscience* 3(1), 1–9.
- Verplaetse, T., Sanfilippo, F., Rutle, A., Osen, O. L. and Bye, R. T. (2016), On Usage of EEG Brain Control for Rehabilitation of Stroke Patients, Proceedings of the 30th European Conference on Modelling and Simulation (ECMS '16) (submitted for publication), Regensburg, Germany.
- Yegnanarayana, B. (2009), *Artificial neural networks*, PHI Learning Pvt. Ltd.
- Zhu, D., Bieger, J., Molina, G. G. and Aarts, R. M. (2010), A survey of stimulation methods used in SSVEP-based BCIs, *Computational intelligence and neuroscience* 2010, 1.

AUTHOR BIOGRAPHIES

ROLF-MAGNUS HJØRUNGDAL is currently a MSc

student in simulation and visualisation at NTNU in Ålesund (formerly Aalesund University College). He completed his BSc degree in automation engineering in 2015, where his thesis provided the foundation for the project described in this paper.

FILIPPO SANFILIPPO¹ received the BSc degree in computer engineering from the University of Catania, Catania, Italy, in 2009 and the MSc degree in computer engineering from the University of Siena, Siena, Italy, in 2011. In 2008, he was a Visiting Scholar at the School of Computing and Intelligent Systems, University of Ulster, Londonderry, United Kingdom and in 2010 a Visiting Fellow at the Technical Aspects of Multimodal Systems (TAMS) research group, Department of Mathematics, Informatics and Natural Sciences, University of Hamburg, Hamburg, Germany. In 2015, he received a PhD degree from the Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. For his PhD studies, he was awarded a research scholarship from the IEEE Oceanic Engineering Society (OES) Scholarship program. He is currently working as a Post-Doctoral Researcher at the Department of Engineering Cybernetics, NTNU in Trondheim, Norway. His research interests include control methods, robotics, artificial intelligence and modular robotic grasping.

OTTAR L. OSEN is MSc in Cybernetics from the Norwegian Institute of Technology in 1991. He is the head of R&D at ICD Software AS and an assistant professor at NTNU in Ålesund.

ADRIAN RUTLE² holds PhD and MSc degrees in Computer Science from the University of Bergen, Norway. Rutle is an associate professor at the Department of Computing, Physics and Mathematics at the Bergen University College, Norway. Rutle's main interest is applying theoretical results from the field of model-driven software engineering to practical domains and has expertise in the development of formal modelling frameworks and domain-specific modelling languages. He also conducts research in the fields of modelling and simulation for virtual prototyping purposes.

ROBIN T. BYE³ 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 NTNU in Ålesund (formerly Aalesund University College) as a researcher in 2008 and has since 2010 been an associate professor in automation engineering. His research interests belong to the fields of artificial intelligence, cybernetics, and neuroengineering.

¹filipposanfilippo.inspitivity.com

²www.rutle.no

³www.robinbye.com

INTELLIGENT COMPUTER-AUTOMATED CRANE DESIGN USING AN ONLINE CRANE PROTOTYPING TOOL

Ibrahim A. Hameed*, Robin T. Bye*, Ottar L. Osen*,†
Birger Skogeng Pedersen*, and Hans Georg Schaathun*

* Software and Intelligent Control Engineering Laboratory
Faculty of Engineering and Natural Sciences
Norwegian University of Science and Technology
NTNU in Ålesund, Postboks 1517, NO-6025 Ålesund, Norway
† ICD Software AS
Hundsværgata 8, NO-6008 Ålesund, Norway

KEYWORDS

Virtual Prototyping; Product Optimisation; Artificial Intelligence; Genetic Algorithm.

ABSTRACT

In an accompanying paper submitted concurrently to this conference, we present our first complete version of a generic and modular software framework for intelligent computer-automated product design. The framework has been implemented with a client-server software architecture that automates the design of offshore cranes. The framework was demonstrated by means of a case study where we used a genetic algorithm (GA) to optimise the crane design of a real and delivered knuckleboom crane. For the chosen objective function, the optimised crane design outperformed the real crane. In this paper, we augment our aforementioned case study by implementing a new crane optimisation client in Matlab that uses a GA both for optimising a set of objective functions and for multi-objective optimisation. Communicating with an online crane prototyping tool, the optimisation client and its GA are able to optimise crane designs with respect to two selected design criteria: the maximum safe working load and the total crane weight. Our work demonstrates the modularity of the software framework as well as the viability of our approach for intelligent computer-automated design, whilst the results are valuable for informing future directions of our research.

INTRODUCTION

The need to reduce the time and cost involved in taking a product from conceptualisation to production and the desire to meet customers' demands and their ability to compete have encouraged companies to turn to new and emerging technologies in the area of manufacturing. One such technology is virtual prototyping (VP) (Mujber et al., 2004). VP refers to the process of simulating the user, the product, and their combined (physical) interaction in software through the different stages of product design, and the quantitative performance analysis of the product

(Song et al., 1999). Being a relatively new technology, VP typically involve the use of virtual reality (VR), virtual environments (VE), computer-automated design (CautoD) solutions, computer-aided design (CAD) tools, and other computer technologies to create digital prototypes (e.g., Gowda et al., 1999).

Together with two companies in the industrial maritime cluster of Norway, ICD Software AS (provider of industrial control systems software)¹ and Seanics AS (designer and manufacturer of offshore equipment)², we have received funding from the Research Council of Norway and its Programme for Regional R&D and Innovation (VRI) for two independent but related research projects (grant nos. 241238 and 249171) for using artificial intelligence (AI) for intelligent computer-automated design (CautoD) of offshore cranes and winches, respectively. In an accompanying paper submitted concurrently (Bye et al., 2016), we present our first complete version of a generic and modular software framework for intelligent computer-automated product design. The framework has been implemented with a client-server software architecture for the design of offshore cranes and consists of several modules: a server-side crane prototyping tool (CPT); a client-side web graphical user interface (GUI); and a client-side artificial intelligence for product optimisation (AIPO) module that uses a genetic algorithm (GA) for optimisation.

The framework was demonstrated by means of a case study where we used the AIPO module and its GA to optimise the crane design of a particular real-world knuckleboom crane that has already been designed by Seanics AS and sold to a company in Baku, Azerbaijan, for a total delivery price of approximately 2.9 million EUR. For the chosen objective function, the optimised crane design outperformed the real crane.

Motivation and Aim

For the work we present here, we will focus solely on intelligent CautoD of offshore cranes, using the software framework developed concurrently (Bye et al., 2016). In

Corresponding author: Ibrahim A. Hameed, ibib@ntnu.no.

¹www.icdsoftware.no

²www.seanics.com

the concurrent paper, we tested the framework with a case study that only involved a single crane optimisation client using a single objective function. Here, we aim at complementing this work by completing the following three goals: (i) examine the modularity of the framework by developing a new crane optimisation client in Matlab (the MCOC module) to be used instead of the AIPO module; (ii) augment the abovementioned case study with a set of alternative objective functions as well as multi-objective optimisation (MOO); and (iii) interpret the results to inform the directions of future work.

To make this paper self-contained, we reproduce some of the material from our accompanying paper (Bye et al., 2016). However, much of the relevant background literature pertaining to VP, CautoD, and design of offshore cranes has been left out. The interested reader is encouraged to read the accompanying paper for further details.

METHOD

This section outlines the software architecture and describes the main components. We provide details on GAs, objective functions, and multi-objective optimisation, before we present a case study on intelligent CautoD of offshore cranes.

Software Architecture

The diagram in Figure 1 shows the client-server software architecture of the framework that we present in our accompanying paper (Bye et al., 2016). On the server-side, the CPT is able to calculate a number of key performance indicators (KPIs) of a specified crane design based on a set of about 120 design parameters. On the client-side, the web GUI facilitates the process of manually selecting the design parameters of the designed CPT and providing a simple visualisation of the designed crane and its 2D workspace safe working load (SWL) chart. Additionally, the AIPO module that uses a GA for optimising the design parameters in a manner that achieves the crane's desired design criteria (that is, the level or quality of the KPIs, typically related to performance and cost).

In the work we present here, we replaced the AIPO module and its GA library with a new Matlab software module that implements a crane optimisation client for CautoD, the MCOC module. To emphasise that the framework is generic and modular, we chose to use the WebSocket (WS) communication interface instead of the hypertext transfer protocol (HTTP) that the AIPO module used (see Figure 1). WS is a protocol providing full-duplex communication channels over a single TCP connection. Because WS enables streams of messages on top of TCP, using WS for communication is advantageous for bidirectional conversations involving many small messages being sent to and from a server. JavaScript Object Notation (JSON), a lightweight human-readable data-interchange format, was used for data messages. We also kept the existing web GUI in order to obtain visualisations of load charts. The software architecture for this reduced subsystem is highlighted with white boxes and solid connections

in Figure 1, whereas the remaining boxes in grey and the dashed lines indicate modules and their interconnections outside the scope of this paper.

Online Crane Prototyping Tool (CPT)

The CPT server consists of a crane calculator and two modules for handling WS/JSON and HTTP/JSON connections (see Figure 1). Here, we let our MCOC connect via WS/JSON to the CPT (see Figure 1). Messages are sent as JSON objects in a standardised format that the CPT accepts, consisting of three parts (subobjects): (i) a "base" object with a complete set of default design parameter values; (ii) a "mods" object with a subset of design parameter values that modifies the corresponding default values; and (iii) a "kpis" object with the desired KPIs to be calculated and returned by the CPT.

Crane Calculator

The components of an offshore crane may total several thousand parameters, making it infeasible to manually pick good values for each parameter. However, through the years, crane designers have been able to reduce this number to a set of about 120 design parameters that are considered the most important. Based on the values of these parameters, which can be set manually or by a CautoD tool such as MCOC, our crane calculator is able to calculate a fully specified crane design and its associated KPIs. The goal of the designer is thus to determine appropriate design parameter values that achieve desired design criteria (based on KPIs), while simultaneously meeting requirements by laws, regulations, codes and standards.

The accuracy of our crane calculator has been verified against other crane calculators and spreadsheets currently in use in the industry, and, as a result, Seaonics AS has already adopted the CPT server and web GUI client for manual crane design.

Web Graphical User Interface (GUI)

To simplify practical use of the crane calculator, we have created a web graphical user interface (GUI) that can be used to interact with the crane calculator via WS/JSON communication. Using the web GUI to manually adjust the 120 design parameters in the crane calculator by trial-and-error, the effect of the parameters on a number of KPIs and other design criteria can be investigated numerically, with the possibility for exporting to text files, and visually, by depicting the main components of the crane and its 2D SWL load chart.

Due to space consideration, we refer to Bye et al. (2015, 2016) for a screenshot of the GUI and more information.

Matlab Crane Optimisation Client (MCOC)

The manual design process using the web GUI together with the CPT is cumbersome. Indeed, there are more than 120 parameters that must be specified by the crane designer. Clearly, this large number of parameters makes

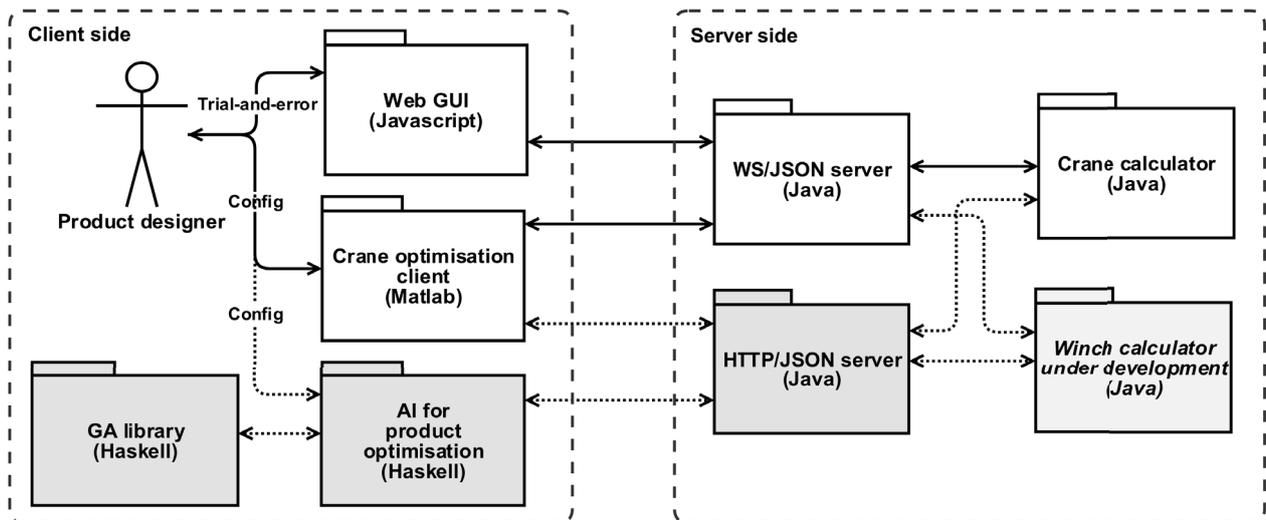


Figure 1: Generic and modular software architecture for intelligent CautoD of offshore cranes, winches, or other products. The modules in white (grey) and their solid (dashed) interconnections are inside (outside) the scope of this paper.

the search space (the space of all possible combinations of parameter values) very large and a manual trial-and-error approach will necessarily be both time-consuming and cost-inefficient and lead to suboptimal designs.

In our accompanying paper (Bye et al., 2016), we present an AIPO software module replacing the human crane designer in order to automate and optimise the design process. Here, we use Matlab to implement such a crane optimisation client, the MCOC module. Two libraries freely available from the MathWorks File Exchange³ were used for the WS/JSON interface, namely MatlabWebSocket, which is a simple library consisting of a websocket server and client for Matlab, and JSONlab, which is a toolbox to encode/decode JSON files in Matlab. For optimisation, we used the GA Solver and the Multiobjective GA Solver from the Global optimisation Toolbox (Mathworks, Inc., 2015). The GA solvers were used to optimise a set of objective functions that we define later.

The Genetic Algorithm (GA)

A GA is a search method based on principles of natural selection and genetics (Holland, 1975). GAs encode the decision variables of a search problem into finite-length strings of alphabets of certain cardinality. The strings, which are candidate solutions to the search problem, are referred to as *chromosomes*, the alphabets are referred to as *genes*, and the values of the genes are called *alleles*. In contrast to traditional optimisation techniques, GAs work with coding of parameters, rather than the parameters themselves. To evolve good solutions and to implement natural selection, a measure for distinguishing good solutions from bad solutions is required. This measure is usually an objective function, and is called a fitness (cost) function if the goal is to maximise (minimise) it.

³<http://www.mathworks.com/matlabcentral/fileexchange>

Another important concept of GAs is the notion of population. Unlike most traditional search methods, GAs rely on a population of candidate solutions. The population size, which is usually a user-specified parameter, is one of the important factors affecting the scalability and performance of GAs. A small population size might lead to premature convergence and yield substandard solutions. On the other hand, large population sizes lead to unnecessary expenditure of valuable computational time. Once the problem is encoded in a chromosomal manner and a fitness or cost measure for discriminating good solutions from bad ones has been chosen, a GA can start to evolve solutions to the search problem using the following steps:

- 1) *Initialization*. The initial population of candidate solutions is usually generated randomly across the search space. However, domain-specific knowledge or other information can be easily incorporated.
- 2) *Evaluation*. Once the population is initialized or an offspring population is created, the fitness values of the candidate solutions are evaluated.
- 3) *Selection*. Selection allocates more copies of those solutions with higher fitness (lower cost) and thus imposes the survival-of-the-fittest mechanism on the candidate solutions. The main idea of selection is to prefer better solutions to worse ones, and many selection procedures have been proposed to accomplish this idea, including roulette-wheel selection, stochastic universal selection, ranking selection and tournament selection.
- 4) *Recombination*. Recombination combines parts of two or more parental solutions to create new, possibly better solutions (i.e. offspring). There are many ways of accomplishing this, and good performance depends on a properly designed recombination mechanism.
- 5) *Mutation*. While recombination operates on two or

more parental chromosomes, mutation locally but randomly modifies a solution.

- 6) *Replacement*. The offspring population created by selection, recombination, and mutation replaces the original parental population. Many replacement techniques such as elitist replacement, generation-wise replacement and steady-state replacement methods are used in GAs.
- 7) *Repeat*. Steps 2–6 are repeated until a termination criterion is satisfied, for example, a maximum number of generations, a run-time limit, a fitness threshold, or no improvement is detected for certain number of generations or run-time.

Objective Functions

In GAs, an objective function (either a cost function or a fitness function) is used to generate an output from a set of input variables (a chromosome). The goal is to modify the output in some desirable fashion by finding the appropriate values for the input variables (Haupt and Haupt, 2004).

GAs are generally customised for solving single-objective optimisation problems (SOPs). However, many, or most, real-world engineering problems require MOO, since they have multiple, often conflicting, objectives such as minimising cost while maximising performance. GAs can be used for MOO through the aggregation of the individual objective functions into a single composite function. Determination of a single objective is possible with methods such as utility theory or the weighted sum method but the problem lies in the correct selection of the weights or utility functions to characterise the decision-makers' criteria. In practice, it can be very difficult to precisely and accurately select these weights, even for someone very familiar with the problem domain. Also, small perturbations in the weights can lead to very different solutions. For this reason and others, decision-makers often prefer a set of promising solutions given the multiple objectives (Konak et al., 2006). Such a set is called a Pareto optimal set of solutions.

Multi-Objective optimisation using a GA (MOOGA)

Combining individual objective functions into a single composite objective function is challenging and might not be realistic or even correct. The second general approach is to determine an entire Pareto optimal solution set or a representative subset. A Pareto optimal set is a set of solutions that are non-dominated with respect to each other. While moving from one Pareto solution to another, there is always a certain amount of sacrifice in one objective to achieve a certain amount of gain in the other. Determining a set of Pareto solutions overcomes the problem of weight selection often used in when combining individual objectives into one composite objective function.

Case Study

We adopt the same case study as in our accompanying paper (Bye et al., 2016), where a real knuckleboom crane

is used as a nominal benchmark against an optimised crane. The crane has about 120 different design parameters and a number of KPIs. Due to the large number of design parameters, the manual design process is cumbersome, time consuming and expensive. Even simple versions of such offshore cranes consist of a large number of components, including hooks, winches, slewing rings, cylinders, booms, hinges, sheaves, and pedestals. Figure 2 illustrates the main components of offshore cranes.

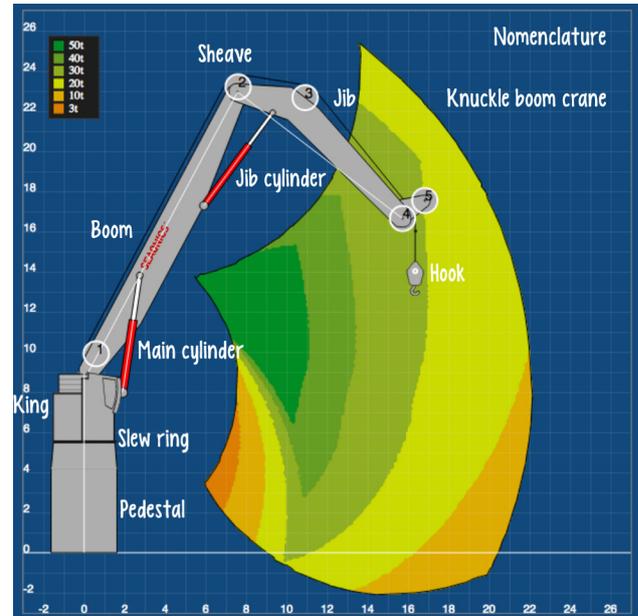


Figure 2: Illustration of the main components of an offshore knuckleboom crane and its 2D load chart. Image courtesy of ICD Software AS.

In an attempt to reduce design time, cost and satisfy customers' need, we propose a CautoD solution in which the MCOG module uses a GA to automate the process and optimise the design.

Choice of KPIs: Among many relevant KPIs, two KPIs were chosen as components of the objective functions in order to demonstrate proof-of-concept, namely the maximum safe working load SWL_{max} and the total crane weight W . Whilst the total crane delivery price is of great concern, currently there is no function implemented in the CPT that can precisely estimate the delivery price of the crane designed. Nevertheless, the total weight can, to some extent, be used as a proxy for price, because price will have some correlation to the weight, and one wants to minimise both measures. Moreover, these cranes are installed on-board vessels and a reduced crane weight allows for a higher deadweight tonnage (DWT). Hence, weight is important for both capital and operating expenditure. The maximum SWL_{max} , on the other hand, is a measure of the maximum safe lifting capacity of the crane within the workspace. The goal of the design is to maximize SWL_{max} while simultaneously minimising W . These two objectives are conflicting and competing with each other, since increasing SWL_{max} will tend to increase W and vice versa.

A number of objective functions were implemented in the MCOC and are presented below.

Objective function f_1 : An intuitive choice for an objective function composed of SWL_{\max} and W is the fitness function f_1 given by

$$f_1 = \frac{SWL_{\max}}{W}, \quad (1)$$

since the evaluation of f_1 will increase when SWL_{\max} increases and/or W decreases, and vice versa.

Objective function f_2 : Another composite objective function f_2 is the weighted sum of both SWL_{\max} and W given by

$$f_2 = w_1 SWL_{\max} + w_2 \frac{1}{W}, \quad (2)$$

where w_1 and w_2 are weight values used to reflect the importance or amount of contribution of SWL_{\max} and W . We note that the total fitness will increase when SWL_{\max} increases and/or W decreases, and vice versa.

Objective functions f_3 and f_4 : It may be of interest to design a crane where either SWL_{\max} or W is the same as for the nominal benchmark crane, while we optimise the remaining KPI. For example, it might be that the crane customer wants a crane with the same “target” weight W_{target} as the nominal crane but with a higher SWL_{\max} . Likewise, the crane customer might require the optimised crane to be able to safely lift as much (but not necessarily more, since this could for example have a detrimental effect on the delivery price) as the benchmark crane, denoted as SWL_{target} , but with a smaller total crane weight W . The objective function must therefore “punish” deviations from the target KPI while optimising the other KPI. Thus, two possible cost functions f_3 and f_4 are given by

$$f_3 = w_1 \frac{1}{SWL_{\max}} + w_2 \left| W_{\text{target}} - W \right| \quad (3)$$

and

$$f_4 = w_1 \left| SWL_{\text{target}} - SWL_{\max} \right| + w_2 W, \quad (4)$$

where w_1 and w_2 are weight values as before.

Choice of Optimisation Variables: Among the 120 different design parameters, four design parameters that greatly affect both SWL_{\max} and W were chosen as decision (i.e., optimisation) variables, namely (i) the boom length L_{boom} ; (ii) the jib length L_{jib} ; (iii) the maximum pressure of the boom cylinder $P_{\text{max,boom}}$; and (iv) the maximum pressure of the jib cylinder $P_{\text{max,jib}}$. The parameter values were constrained to a range with minimum and maximum limits. All other design parameters were identical to those of the nominal crane.

GA Settings: For GA optimisation, we used a population size (set of candidate design solutions) of 100 and let the GA run for 50 generations, giving a grand total of 5,000 evaluated designs.

RESULTS

Table 1 shows the values of the four design parameters L_{boom} , L_{jib} , $P_{\text{max,boom}}$, and $P_{\text{max,jib}}$ and the resulting maximum SWL (SWL_{\max}) and total crane weight (W) for the nominal crane that we use as a benchmark with which to compare the optimisation results. During optimisation, each design parameter was constrained to a minimum and a maximum value as given by the Table 1. The table also shows the objective function evaluations of the nominal crane.

measure	units	nominal	(min, max)
L_{boom}	mm	15800	(12000, 26000)
L_{jib}	mm	10300	(6000, 16000)
$P_{\text{max,boom}}$	bar	315	(100, 400)
$P_{\text{max,jib}}$	bar	215	(50, 300)
SWL_{\max}	tonne	99.978	-
W	tonne	50.856	-
objective function	evaluation	w_1	w_2
f_1	1.9659	-	-
f_2	100.00	1	1
f_2	198.29	1	5000
f_2	1098.10	10	5000
f_2	108.31	0.1	5000
f_3	0.01000	1	1
f_4	50.856	1	1

Table 1: Nominal crane, its objective function evaluations, and optimisation constraints.

Table 2 provides a summary of the results. It shows the total processing times and optimised values for SWL_{\max} and W for each of the optimised cranes, the mean and standard deviation for these values, and the difference of the means when compared with the nominal crane.

objective function	SWL_{\max}	W	T (min)
f_1	142.14	44.01	98.4
$f_2, w_1 = w_2 = 1$	140.63	44.22	115.21
$f_2, w_1 = 1, w_2 = 5000$	140.59	44.22	89.39
$f_2, w_1 = 10, w_2 = 5000$	140.02	44.22	106.19
$f_2, w_1 = 0.1, w_2 = 5000$	143.37	43.88	66.36
$f_3, w_1 = w_2 = 1$	112.54	50.81	125.82
$f_4, w_1 = w_2 = 1$	99.94	47.1	90.97
MOO	140.95	43.88	182.83
mean	132.52	45.29	109.40
standard deviation	16.60	2.47	34.68
nominal	99.98	50.86	-
difference of mean with nominal	32.54	-5.56	-

Table 2: Processing time T in minutes and optimal values of SWL_{\max} and W for the set of objective functions, their mean and standard deviation, and the difference of the means from the nominal crane.

The total processing time is the total run-time from the start of the optimisation process till a result was obtained, including transfer times between the MCOC client and the CPT server.

For reference, we include the detailed results of employing f_1 – f_4 and MOO for optimisation in Tables 4–11.

Maximum SWL (SWL_{max})

Table 2 shows that employing f_1 , f_2 , or MOO all resulted in optimised cranes with a SWL_{max} greater than 140 tonnes, or an improvement of more than 40 tonnes when compared to the $SWL_{max} = 99.98$ tonnes of the nominal crane.

Employing f_3 , whose purpose is to maximise SWL_{max} while having a W as close as possible to that of the nominal crane, resulted in an SWL_{max} of about 12 tonnes more than the nominal crane's SWL_{max} .

Finally, employing f_4 resulted in a crane with a $SWL_{max} = 99.94$, which is almost identical to the $SWL_{max} = 99.98$ tonnes of the nominal crane. This is not surprising, given that the purpose of f_4 was to minimise W while having a SWL_{max} as close as possible to that of the nominal crane.

The mean SWL_{max} for all the optimised crane designs was 132.52 tonnes, or an improvement of 32.54 tonnes when compared with the nominal crane. The standard deviation of SWL_{max} for the optimised cranes was 16.60.

Total Crane Weight (W)

Table 2 shows that employing f_1 , f_2 , or MOO all resulted in optimised cranes with a total W of 44.22 tonnes or less, or an improvement of about 7 tonnes when compared to the $W = 50.86$ tonnes of the nominal crane.

Employing f_4 resulted in a $SWL_{max} = 47.1$, or an improvement of nearly 4 tonnes when compared to the nominal crane.

Finally, employing f_3 resulted in a crane with a $W = 50.81$, which is almost identical to that of the nominal crane.

The mean W for all the optimised crane designs was 45.29 tonnes, or an improvement of 5.56 tonnes when compared with the nominal crane. The standard deviation of W for the optimised cranes was 2.47.

Processing Times

The total processing time for each of the optimisation processes ranged from 66.36 minutes for f_2 with $w_1 = 0.1$ and $w_2 = 5000$ to 182.83 minutes for the MOO. The mean processing time was 109.40 minutes, with a standard deviation of 34.68 minutes. The fastest processing time was more than one standard deviation lower than the mean, whereas the slowest processing time was more than two standard deviations higher than the mean. The remaining processing times were all within one standard deviation from the mean.

SWL Load Charts

The SWL load charts for nominal crane and the optimised crane designs are shown in Figures 3–6.

Each load chart shows the workspace and the SWL lifting capacity in various coloured zones of the workspace for a given crane. The legend at the top left indicates the capacity of a particular zone with colours in a spectrum from red (3 tonnes) up to blue (150 tonnes).

Comparing the charts, it is apparent that all the optimised cranes have one or several zones with a SWL capacity in the range 50–150 tonnes, whereas the zone of the nominal crane with the highest SWL capacity is 50 tonnes.

It can also be observed that the overall lifting capacity of the workspace is higher than that of the nominal crane.

However, a notable observation is that all crane designs apart from that obtained using f_3 has a smaller workspace than the nominal crane. The reason for this is that whereas the goal of using f_3 is to obtain a total crane weight W identical to the nominal crane (while maximising SWL_{max}), the other objective functions and the MOO try to minimise W . As a result, the lengths of the boom and jib are shorter than the nominal crane for these latter designs, thus making W smaller, but at the expense of a smaller workspace.

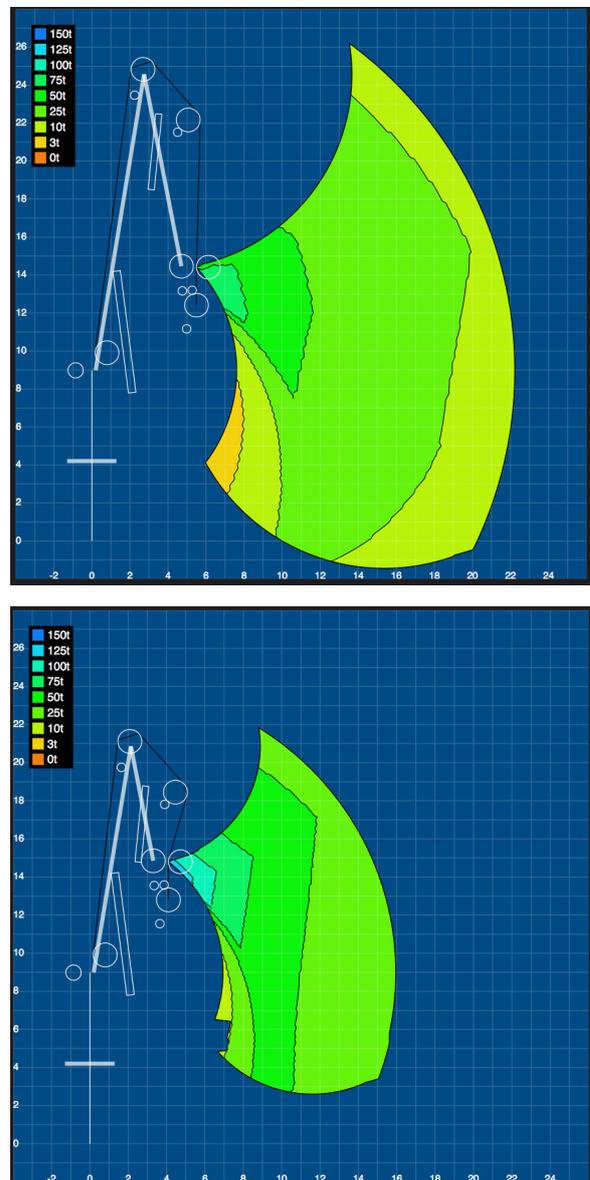


Figure 3: SWL load charts: nominal (top); f_1 (bottom).

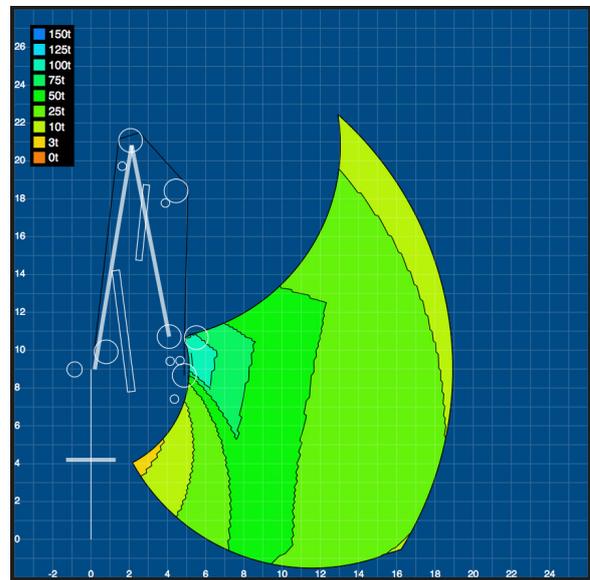
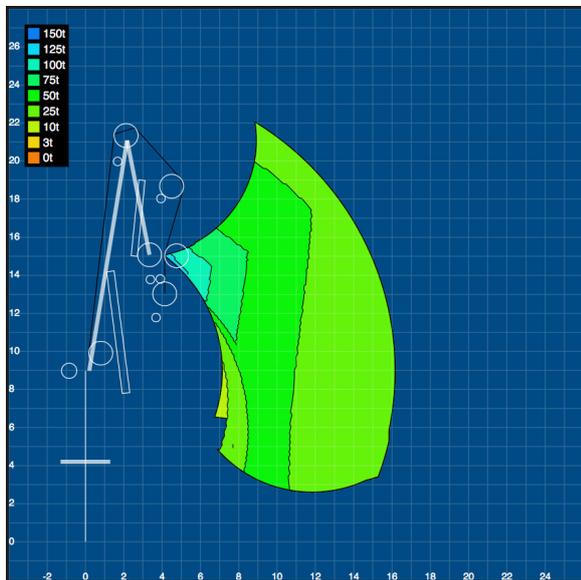
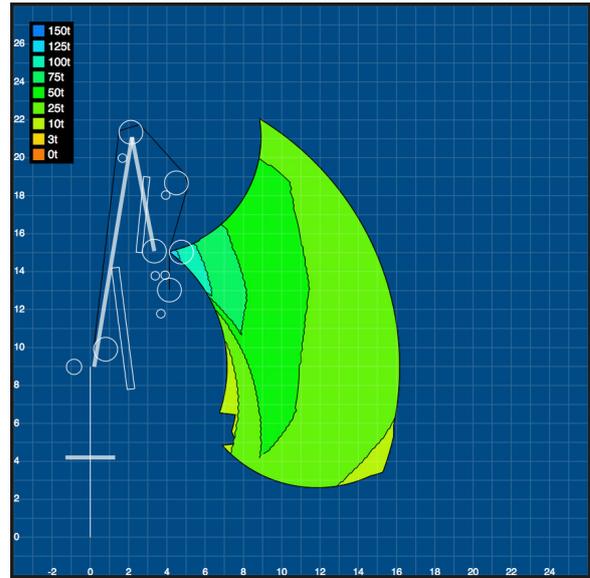
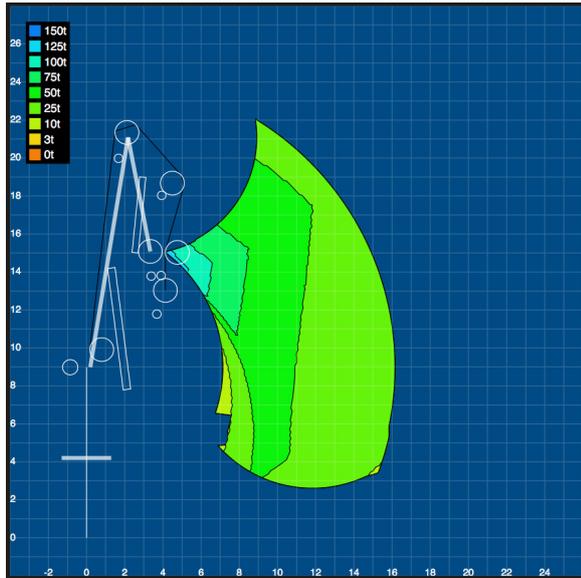


Figure 4: SWL load charts: f_2 : $w_1 = w_2 = 1$ (top); f_2 : $w_1 = 1, w_2 = 5000$ (bottom).

Figure 5: SWL load charts: f_2 : $w_1 = 10, w_2 = 5000$ (top); f_2 : $w_1 = 0.1, w_2 = 5000$ (bottom).

Multiobjective Optimisation (MOO)

For MOO, the two KPIs SWL_{max} and W were used as two individual objective functions to be respectively maximised and minimised by the Matlab MOOGA Solver. The optimal solution is provided as a set of Pareto-optimal solutions for values of the design parameters given by Table 3. Each of these solutions results in a crane design with $SWL_{max} = 140.95$ tonnes and $W = 43.88$ tonnes. A sample solution is presented in Table 11 in the Appendix.

DISCUSSION

In this paper, we have presented an intelligent computer-automated design solution for optimising offshore cranes using a genetic algorithm for single-objective or multi-objective optimisation. Candidate crane designs suggested by a GA incorporated in a Matlab crane optimisation

client are sent to an online crane prototyping tool that uses a crane calculator to determine two key performance indicators, the maximum safe working load and the total crane weight, for each crane design. The CPT server sends the results back to the MCOC and the GA uses them to evolve another set of candidate solutions. The process iterates until some stopping criteria is satisfied, for example when the solutions do not improve for a prolonged number of iterations.

Case Study

To test the viability of our approach, we adopted the case study of our accompanying paper (Bye et al., 2016). Here, we used MOO and a set of four different objective functions to optimise the design of an offshore crane consisting of about 120 design parameters. Of about 120 design parameters, most were fixed to values correspond-

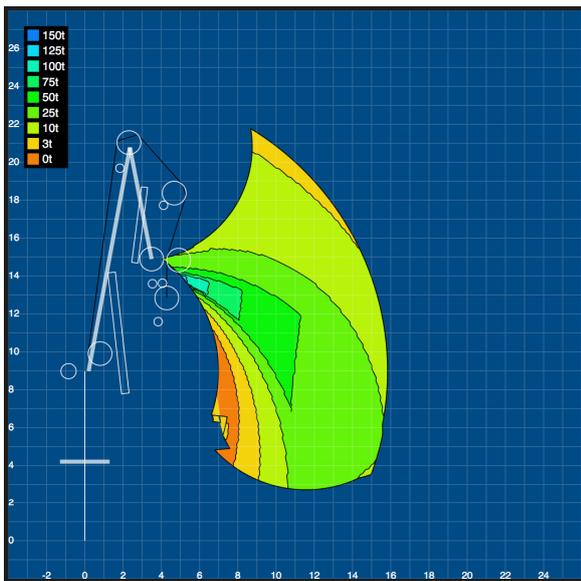
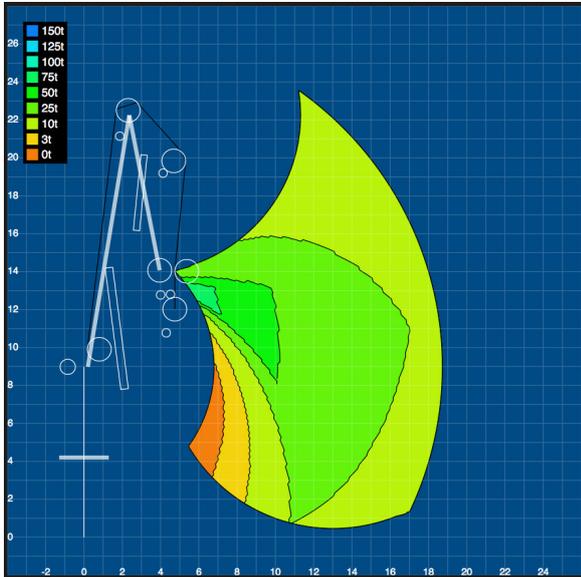
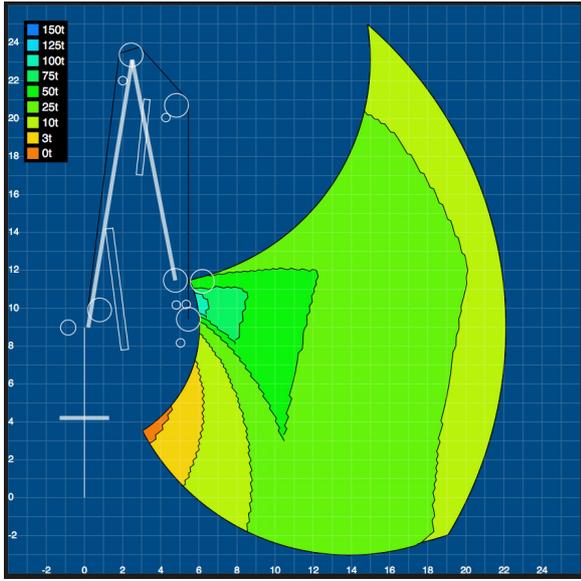


Figure 6: SWL load charts: f_3 (top); f_4 (middle); MOO (bottom).

L_{boom}	L_{jib}	$P_{\text{max,boom}}$	$P_{\text{max,jib}}$
12000.9	6000.95	331.309	67.1085
12000.8	6000.50	331.183	75.4687
12000.9	6000.39	332.326	88.6125
12000.9	6000.60	332.32	71.5644
12000.6	6000.95	331.309	67.1085
12000.6	6000.49	332.241	80.8655
12001.0	6000.69	332.101	72.0046
12000.9	6000.88	331.984	68.08
12000.9	6000.61	332.333	83.6164
12000.9	6000.98	331.788	74.0161
12001.0	6000.99	331.876	75.4702
12000.8	6000.95	331.309	67.1554
12000.9	6000.74	332.816	66.7924
12000.9	6000.55	331.308	75.4687
12000.7	6000.92	332.344	82.1006
12000.8	6000.61	332.227	82.379
12000.8	6000.82	332.414	66.8399
12000.5	6000.97	331.109	72.9711
12000.7	6000.79	331.142	87.5899
12000.6	6000.99	332.633	81.1601
12000.6	6000.95	331.151	84.4106
12000.8	6000.93	332.252	89.7784
12000.9	6000.86	331.048	82.0198
12000.6	6000.81	331.017	80.539
12000.9	6000.51	332.668	68.4269
12000.7	6000.98	331.41	79.5427
12000.9	6000.91	331.983	68.8521
12000.7	6000.99	331.75	81.9171
12000.7	6000.66	332.401	82.9699
12000.8	6000.82	331.325	72.0107
12000.9	6000.95	331.246	67.1085
12000.7	6000.49	332.264	80.928
12000.9	6000.93	331.261	68.3642
12000.8	6000.63	332.088	78.7431
12000.9	6000.51	332.668	68.4269

Table 3: Pareto set of optimised cranes using MOO that all have $SWL_{\text{max}} = 140.95$ and $W = 43.88$.

ing to the design of a real and delivered crane, whereas four design parameters, the boom length, jib length, maximum boom cylinder pressure, and maximum jib cylinder pressure, were optimised by the MCOC. The goal of the optimisation was to maximise the lifting capacity given by SWL_{max} and the total crane weight W . Two of the objective functions (f_3 and f_4) only tried to optimise one of the KPIs while keeping the other as close as possible to that of the nominal crane.

The results show that all the optimised crane designs outperform the nominal crane on the two selected KPIs. However, other KPIs not incorporated in the optimisation process will inevitably also be change when the crane design changes. This can lead to unwanted and unexpected results. For example, as can be seen from the load charts in Figures 3–6, whilst the optimised crane designs have improved the maximum SWL, most of them have reduced workspace as a sideeffect. One way to overcome this would be to incorporate another KPI, or optimisation objective, relating to the workspace area.

Future Work

Importantly, this case study was limited to optimising only a fraction of all the design parameters needed to construct an offshore crane, and only two KPIs were considered. For more realistic use, our method needs to

be expanded to involve both more design parameters and more KPIs. The first is trivial, as it only involves minor modification to the GA; the latter is non-trivial, as many KPIs are interrelated and mutually conflicting (for example, delivery cost versus performance), and care must be taken in the choice of objective functions. We plan to work in close cooperation with our industrial partners and their crane designers to develop a set of useful KPIs and objective functions for real-world use.

Using optimisation weights for single-objective optimisation is one means for handling this problem, however, choosing the right weights can be difficult. Using MOO can, at least to some extent, handle the problem automatically without the need to determine such weights. Instead of a single design solution, one obtains a Pareto set of solutions, all with the same values for the desired KPIs. The crane designer and customer must then decide which solution in the set to implement and build for delivery.

Being able to handle many more design parameters and KPIs will likely lead to slower processing times, since objective function evaluation is the main contributor to computational load. Still, for the proof-of-concept study we do here, the mean processing time was less than 2 hours, which is many orders of magnitude smaller than what a human crane designer would require. In future work, we intend to implement several other AI algorithms for optimisation, possibly using parallel computation, and examine their performance both with respect to optimisation and processing time.

We will also use the knowledge we gain from our study on offshore cranes to inform related work on optimised CautOD of offshore winches.

Concluding Remarks

The work presented here has accomplished the three goals set out in the introduction: We have successfully been able to use the software framework developed concurrently (see Bye et al., 2016) by creating a new product optimisation client customised for offshore cranes and insert it as a module in our existing framework. Moreover, we have augmented the case study we present in Bye et al. (2016) with a set of alternative objective functions and with MOO and the results are valuable for our future development. Finally, we would like to point the reader to our accompanying paper for related concurrent and future work (Bye et al., 2016).

ACKNOWLEDGEMENTS

The SoftICE lab at NTNU in Ålesund wishes to thank ICD Software AS for their contribution towards the implementation of the simulator, and Seonics AS for providing documentation and insight into the design and manufacturing process of offshore cranes. We are also grateful for the support provided by Regionalt Forskningsfond (RFF) Midt-Norge and the Research Council of Norway through the VRI research projects *Artificial Intelligence for Crane Design (Kunstig intelligens for krandesign (KIK))*, grant no. 241238 and *Artificial Intelligence for*

Winch Design (Kunstig intelligens for vinsjdesign (KIV)), grant no. 249171.

REFERENCES

- Bye, R. T., Osen, O. L. and Pedersen, B. S. (2015), A computer-automated design tool for intelligent virtual prototyping of offshore cranes, Proceedings of the 29th European Conference on Modelling and Simulation (ECMS'15), pp. 147–156.
- Bye, R. T., Osen, O. L., Pedersen, B. S., Hameed, I. A. and Schaathun, H. G. (2016), A software framework for intelligent computer-automated product design, Proceedings of the 30th European Conference on Modelling and Simulation (ECMS'16) (submitted for publication).
- Gowda, S., Jayaram, S. and Jayaram, U. (1999), Architectures for internet-based collaborative virtual prototyping, the 1999 ASME Design Technical Conference and Computers in Engineering Conference, DETC99/CIE-9040, Las Vegas, Nevada.
- Haupt, R. L. and Haupt, S. E. (2004), *Practical Genetic Algorithms*, John Wiley & Sons, Inc.
- Holland, J. H. (1975), *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*, University of Michigan Press, Oxford, England.
- Konak, A., Coit, D. W. and Smith, A. E. (2006), Multi-objective optimization using genetic algorithms: A tutorial., *Reliability Engineering & System Safety* 91(9), 992–1007. Springer-Verlag: Berlin Heidelberg.
- Mathworks, Inc. (2015), *MATLAB Global Optimization Toolbox, R2015b*, The Mathworks, Inc., Natick, Massachusetts.
- Mujber, T., Szecsi, T. and Hashmi, M. (2004), Virtual reality applications in manufacturing process simulation, *Journal of Materials Processing Technology* 155, 1834–1838.
- Song, P., Krovi, V., Kumar, V. and Mahoney, R. (1999), Design and virtual prototyping of humanworn manipulation devices, the 1999 ASME Design Technical Conference and Computers in Engineering Conference, DETC99/CIE-9029, Las Vegas, Nevada.

AUTHOR BIOGRAPHIES

IBRAHIM A. HAMEED has a BSc and a MSc in Industrial Electronics and Control Engineering, Menofia University, Egypt, a PhD in Industrial Systems and Information Engineering from Korea University, S. Korea, and a PhD in Mechanical Engineering, Aarhus University, Denmark. He has been working as an associate Professor at NTNU in Ålesund since 2015. His research interests include artificial Intelligence, optimization, control systems and robotics.

ROBIN T. BYE⁴ 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 NTNU in Ålesund (formerly Aalesund University College) as a researcher in 2008 and has since 2010 been an associate professor in automation engineering. His research interests belong to the fields of artificial intelligence, cybernetics, and neuroengineering.

OTTAR L. OSEN is MSc in Cybernetics from the Norwegian Institute of Technology in 1991. He is the head of R&D at ICD Software AS and an assistant professor at NTNU in Ålesund.

BIRGER SKOGENG PEDERSEN graduated from NTNU in Ålesund with a BE in automation engineering

⁴www.robinbye.com

and is a former employee at ICD Software AS, during which time he worked on the research projects described in this paper. He is currently a MSc student of simulation and visualisation as well as a employed as a project manager in maritime technology at NTNU in Ålesund.

HANS GEORG SCHAATHUN Hans Georg Schaathun graduated from the University of Bergen with cand.mag. in 1996 (Mathematics, Economics, and Informatics), cand.scient. in 1999 (Industrial and Applied Mathematics and Informatics), and dr.scient. in 2002 (Informatics – Coding Theory), all from the University of Bergen (UiB), Norway. After a period as a lecturer and a postdoc at UiB and a lecturer and senior lecturer at the University of Surrey, England, he joined Aalesund University College (now NTNU in Ålesund) and became a professor in 2011. His current research focus is software engineering and pedagogy.

APPENDIX

All tables show the optimised values for the four design parameters, the resulting SWL_{max} and W , and the optimised objective function value. Values are compared with the nominal crane and the difference and percentage change is shown.

measure	units	optimised	difference	change
L_{boom}	mm	12038	-3762	-23.81%
L_{jib}	mm	6124	-4176	-40.54%
$P_{max,boom}$	bar	383	68	21.59%
$P_{max,jib}$	bar	262	47	21.86%
SWL_{max}	tonne	142.14	42.16	42.17%
W	tonne	44.01	-6.84	-13.45%
f_1	-	3.23	1.26	64.27%

Table 4: Optimised crane using f_1 .

measure	units	optimised	difference	change
L_{boom}	mm	12266.9	-3533.10	-22.36%
L_{jib}	mm	6120.56	-4179.44	-40.58%
$P_{max,boom}$	bar	340.9	25.90	8.22%
$P_{max,jib}$	bar	266.84	51.84	24.11%
SWL_{max}	tonne	140.63	40.65	40.66%
W	tonne	44.22	-6.64	-13.05%
$f_2, w_1 = w_2 = 1$	-	140.65	40.65	40.66%

Table 5: Optimised crane using $f_2, w_1 = w_2 = 1$.

measure	units	optimised	difference	change
L_{boom}	mm	12266.9	-3533.10	-22.36%
L_{jib}	mm	6123.56	-4176.44	-40.55%
$P_{max,boom}$	bar	392.14	77.14	24.49%
$P_{max,jib}$	bar	297.62	82.62	38.43%
SWL_{max}	tonne	140.59	40.61	40.62%
W	tonne	44.22	-6.64	-13.05%
$f_2, w_1 = 1, w_2 = 5000$	-	253.66	55.37	27.92%

Table 6: Optimised crane using $f_2, w_1 = 1, w_2 = 5000$.

measure	units	optimised	difference	change
L_{boom}	mm	12269.6	-3530.40	-22.34%
L_{jib}	mm	6121.56	-4178.44	-40.57%
$P_{max,boom}$	bar	304.72	-10.28	-3.26%
$P_{max,jib}$	bar	249.9	34.90	16.23%
SWL_{max}	tonne	140.02	40.04	40.05%
W	tonne	44.22	-6.64	-13.05%
$f_2, w_1 = 10, w_2 = 5000$	-	1513.27	415.17	37.81%

Table 7: Optimised crane using $f_2, w_1 = 10, w_2 = 5000$.

measure	units	optimised	difference	change
L_{boom}	mm	12000	-3800.00	-24.05%
L_{jib}	mm	6000	-4300.00	-41.75%
$P_{max,boom}$	bar	353.42	38.42	12.20%
$P_{max,jib}$	bar	297.48	82.48	38.36%
SWL_{max}	tonne	143.37	43.39	43.40%
W	tonne	43.88	-6.98	-13.72%
$f_2, w_1 = 0.1, w_2 = 5000$	-	128.28	19.97	18.44%

Table 8: Optimised crane using $f_2, w_1 = 0.1, w_2 = 5000$.

measure	units	optimised	difference	change
L_{boom}	mm	14321.5	-1478.5	-9.36%
L_{jib}	mm	11864.1	1564.1	15.19%
$P_{max,boom}$	bar	396.89	81.89	26.00%
$P_{max,jib}$	bar	187.75	-27.25	-12.67%
SWL_{max}	tonne	112.54	12.562	12.56%
W	tonne	50.81	-0.046	-0.09%
$f_3, w_1 = w_2 = 1$	-	0.0549	0.0449	448.74%

Table 9: Optimised crane using $f_3, w_1 = w_2 = 1$.

measure	units	optimised	difference	change
L_{boom}	mm	13443.9	-2356.1	-14.91%
L_{jib}	mm	8328.99	-1971.01	-19.14%
$P_{max,boom}$	bar	273.36	-41.64	-13.22%
$P_{max,jib}$	bar	101.05	-113.95	-53.00%
SWL_{max}	tonne	99.94	-0.038	-0.04%
W	tonne	47.1	-3.756	-7.39%
$f_4, w_1 = w_2 = 1$	-	47.138	-3.718	-7.31%

Table 10: Optimised crane using $f_4, w_1 = w_2 = 1$.

measure	units	optimised	difference	change
L_{boom}	mm	12000.9	-3799.1	-24.04%
L_{jib}	mm	6000.95	-4299.05	-41.74%
$P_{max,boom}$	bar	331.31	16.31	5.18%
$P_{max,jib}$	bar	67.11	-147.89	-68.79%
SWL_{max}	tonne	140.95	40.972	40.98%
W	tonne	43.88	-6.976	-13.72%

Table 11: Optimised crane using MOO.

High Performance Modelling and Simulation

Modelling and Simulation of Data Intensive Systems

-

Special Session

Big Data as a Service for Monitoring Cyber-Physical Production Systems

Alessandro Marini, Devis Bianchini
Dept. of Information Engineering
University of Brescia
Brescia, Italy

Email: a.marini011@unibs.it, devis.bianchini@unibs.it

KEYWORDS

Big Data, Data Service, Data as a Service, Internet of Services, Cyber-Physical Production Systems, Product-Service Systems, Manufacturing 4.0, Industry 4.0

ABSTRACT

The introduction of Internet of Services technologies is promoting manufacturing servitization of Cyber Physical Production Systems for the most important Manufacturing 4.0 capabilities, namely self-awareness, self-configuration and self-repairing. In addition, industrial data are emerging as a new industrial asset, creating new opportunities for operations improvement, and increase industrial value through the capitalisation of immaterial assets. These recent research trends also raised several challenges and, among them, Big Data acquisition and storage. In this paper, we describe a Data as a Service approach, designed to deal with the Big Data environment. The service is able to manage data volume and velocity during the data collection phase, accumulating and summarizing measures from the machine fleet, and to properly organize them in order to serve advanced Manufacturing 4.0 facilities. Experiments on service performances demonstrate the efficiency of the proposed service.

I. INTRODUCTION

Recently, Cyber-Physical Production Systems (CPPS) have attracted an ever growing attention to realize advanced Manufacturing 4.0 capabilities, namely self-awareness, self-configuration and self-repairing. Cyber-Physical Production Systems have been defined as novel transformative technologies for managing interconnected systems between their physical assets and computational capabilities [9].

In parallel, the introduction of Internet of Services (IoS) technologies contributed to promote *manufacturing servitization*, defined as the strategic innovation of organisations' capabilities and processes to shift from selling products to selling an integrated product and service offering, i.e., a Product-Service System (PSS) [8]. Service-oriented computing enables modern Manufacturing 4.0 infrastructure to deliver self-awareness, self-configuration and self-repairing as a service, through cloud-based data collection and sharing [6], and adherence to widespread standards, such as XML, JSON and recommendations specifically designed for the Internet of Things (IoT), such as MTCConnect¹. To this aim, advanced

¹MTCConnect Institute, <http://www.mtconnect.org/>.

PSS must be rooted on the collection and organization of huge amounts of data from sensors and machine controls, in order to apply advanced functionalities, such as data mining and analysis algorithm, as well as innovative data visualisation for effective decision support, only to name a few. These data overwhelming issues have been accentuated by the emerging and diffusion of IoT and advanced sensing technologies. Therefore, the actual processing of big data is a key factor for the success of Manufacturing 4.0 goals.

A. Open challenges

As underlined in [5], Manufacturing 4.0 advanced capabilities start from machine self-awareness, that enables CPPS to perform automatic state detection and health assessment. This assumption and the panorama depicted above raised interesting research challenges, that we summarize in the following.

Big data acquisition and storage. Collection of huge amounts of data from the sensors and machine controls is a crucial task that has an impact on all the other 4.0 goals we introduced above. Data are provided at high rate, they present a poor structure (they are basically schemaless) and high variety (e.g., missing or incomplete data) and must be stored and indexed considering the functionalities that will be provided on top of the data infrastructure. This challenge is strictly related to the other ones.

Widespread data relevance. All measures coming from monitored CPPS might have a relevance. Therefore, approaches that only consider measures that are close to critical working conditions of machines, according to FMECA analysis, are not always feasible. Depending on the particular conditions in which a machine is working, anomalous behaviours could be identified, although the machine is not close to breakdown limits. Moreover, unknown working conditions might be detected. All these requirements imply that nearly all data collected from the machine should be properly stored.

Data stream elaboration. Measures gathered from machines are featured by a continuous incoming flow (*data stream*). This has a crucial impact on data acquisition, but also on data elaboration and usage. Data storage, although performed by using performant big data technologies, is not straightforward and should be performed in an efficient way, having in mind the use of such data that will be made by Manufacturing 4.0 applications.

Machine similarity identification. Recent approaches [5], [12] suggest to look for similar assets working in similar

conditions, to apply self-configuration and self-repairing facilities to a machine. Nevertheless, detecting the right factors that influence the machine working conditions is not straightforward. In fact, not all influencing factors are completely observable. Analysis of all measures from the machine fleet, to be combined with other information, such as the operating environment or the chronology of maintenance tasks performed on the machines, could help. Big data collection and management is a key enabler in this sense.

In the following, we will discuss a big data infrastructure, and a data service to access it, that are able to meet the above mentioned challenges. In the conclusions, we will provide some additional considerations about the challenges.

B. Paper contribution

In this paper, we propose a Data as a Service approach, rooted within the Big Data environment. In particular, the implemented service collects measures coming from the machine fleet (through sensors and machine controls) and organizes gathered data in a proper infrastructure according to different multi-dimensional feature spaces. Feature spaces are defined as groups of measured features (e.g., voltage, pressure, acceleration) and are configured to monitor specific aspects during machine operation (e.g., energy consumption, parts subject to wear and tear). Engaged Big Data technologies enable the service to manage data volume and velocity during the data acquisition phase, while proper data organization also speeds up data querying phase, as demonstrated by experimental results on service performances. The service is able to accumulate and summarise knowledge on machine behaviour in various working conditions in order to elaborate information that evolves over time. This in the near future will be used to implement advanced Manufacturing 4.0 facilities (self-awareness, self-configuration and self-repairing).

The paper is organized as follows: Section II introduces the research background; in Section III and IV we present the data infrastructure and the implemented service, respectively; experimental results are shown and discussed in Section V; related work are described in Section VI; Section VII closes the paper with some final remarks.

II. RESEARCH BACKGROUND

Recently, a unified architecture for implementation of CPS has been proposed in [5]. We rely on this vision as research background for our Big Data as a Service approach. In particular, such an architecture is organized over five level, as shown in Figure 1:

- *smart connection level*, focused on data measurement and acquisition from machines through proper sensors and controllers, or from enterprise manufacturing systems (e.g., CMMS - Computerized Maintenance Management Systems); at this level, challenges related to various types of data, that might be schemaless and incomplete, and to the engagement of new IoT technologies such as MTConnect are highlighted; data variety is one of the key aspects characterizing Big Data, for which we propose specific technological solutions in our service;

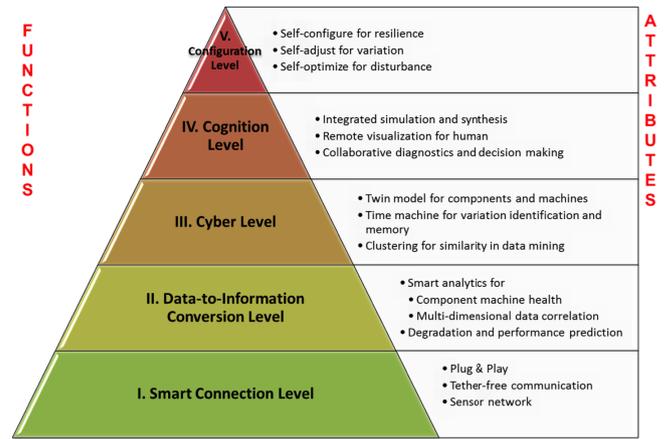


Fig. 1. Five-level architecture for CPS implementation [5].

- *data-to-information conversion level*, focused on tools, methodologies and algorithms specifically conceived to extract meaningful information from gathered data, e.g., implementing advanced prognostics and health assessment applications; this level has been designed to pursue self-awareness in Manufacturing 4.0; our proposed data infrastructure includes information at different granularity levels (e.g., limits detected for different working conditions) in order to enable the application of such tools;
- *Cyber level*, focused on the description in the cyber space of the machine behaviour, in order to reach the capability of machine self-comparison, where performances of a single machine are analyzed in the context of a machine fleet; this aspect is deeply rooted into the concept of similarity between machines, that is one of the challenges highlighted in the introduction;
- *cognition level*, focused on tools and techniques (e.g., info-graphics facilities) to properly present the acquired knowledge to expert users, thus supporting correct decisions to be taken;
- *configuration level*, defined as a feedback from cyber space to the physical one and acting as a resilience control system (RCS) to make machines self-configurable and self-adaptive.

Our proposed Big Data as a Service approach aims at defining a proper data infrastructure that is built on top of the Smart Connection level and is designed to enable the other four levels. Future research that will start from the current version of the service will be discussed in Section VII.

III. THE DATA MODEL

The conceptual model of the data infrastructure underlying our approach is shown in Figure 2. In the following, we will highlight the main elements of the model and the rationale behind their introduction with reference to research challenges.

Hierarchical aggregation of monitored components. Monitored components on the *physical side* (denoted as *Assets*) are hierarchically aggregated as shown in Figure 2, where a *Composite Asset* is defined as an aggregation of

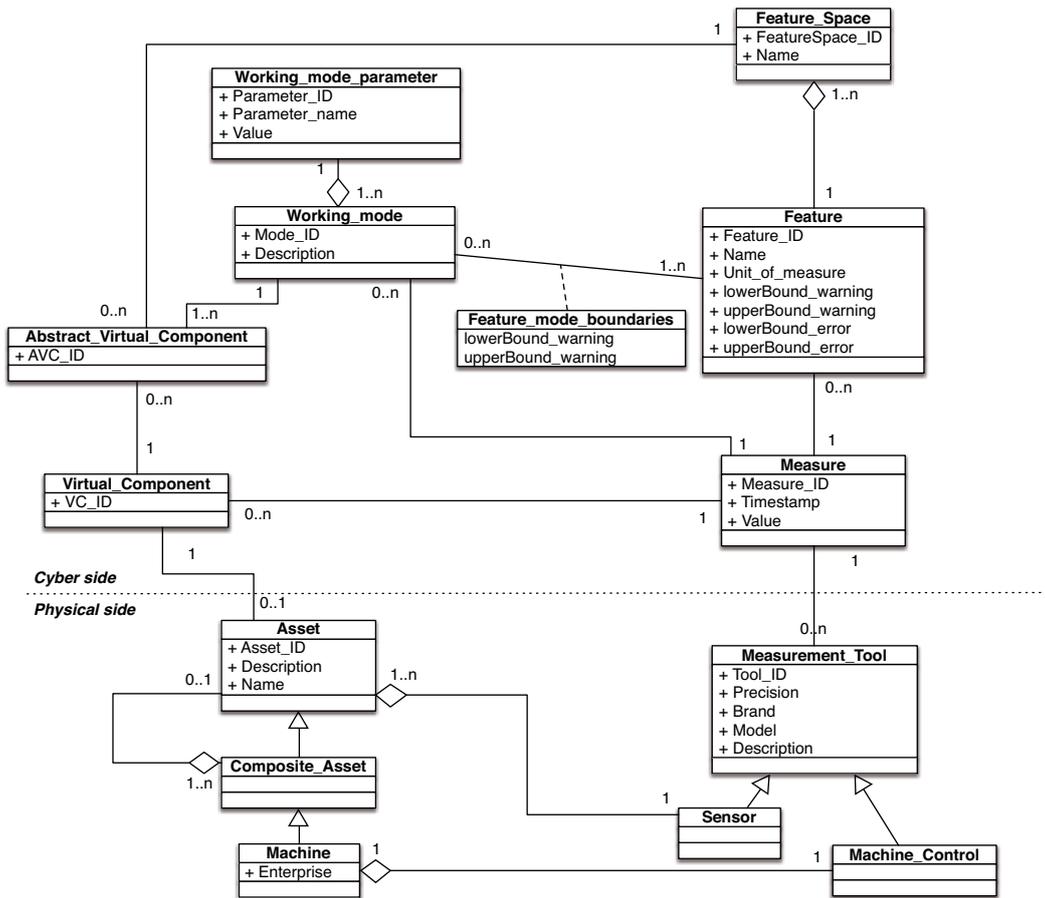


Fig. 2. Conceptual model for the data infrastructure underlying the Big Data as a Service approach.

Assets, that can be either atomic or in turn composite. In this way, the model ensures aggregations of Assets into Composite Assets at arbitrary depth. Among the composite assets, we distinguish the monitored Machine, that constitutes the physical component at the highest level of aggregation. Hierarchical aggregation of physical components can enable the easy propagation of warning/error alerts, that are presented below.

Virtual components and abstract virtual components. The counterpart of a physical component (asset) on the cyber-side is the Virtual Component. Measures always refer to a virtual component, that in turn represents a specific instance of an asset. The Abstract Virtual Component represents a set of virtual components that are mutually similar, that is, act in comparable working conditions, such that boundaries set for state detection alerts are common to all the virtual components associated with the same abstract virtual component. The automatic identification of virtual component similarity involves several (also not observable) factors (e.g., working conditions, contextual variables, maintenance tasks history, and so on) and will be investigated in future work. The conceptual model shown in Figure 2 already takes into account this concept.

Working mode and parameters. All virtual components associated with the same abstract virtual component may operate in different Working modes (but given a working mode, all virtual components associated with the same abstract virtual

component operate coherently by definition). The working mode, for a CPPS, might depend on different Working parameters, such as the task that is being executed (described within the part program), the tool that is being used, the values of specific working variables that have been set through the machine control before starting the task execution. Therefore, monitoring virtual components, associated with the same abstract virtual component, should be performed taking into account the particular working mode in which they are operating.

Features and feature spaces. Features are monitored variables measured through sensors and machine control. Each measure always refer to a specific feature (e.g., voltage, pressure, acceleration), that in turn is described by a name and a unit of measure. Features are aggregated into Feature Spaces. Examples of feature spaces are the machine energy consumption, the degree of wear of a specific asset, and so on. For instance, energy consumption might be measured by considering the voltage, the consumed power, and so on. The goal is to create feature spaces using the different measures needed to describe the evolution in the time of the behaviour of specific machine characteristics. With this multi-dimensional model, we should be able to follow the evolution of the specific monitored characteristics. Multiple feature spaces might be observed, and the observation of a feature might be useful to monitor more than one feature space.

TABLE I. WARNING AND ERROR ALERTS, WITH CORRESPONDING CONTEXT (FEATURE, WORKING MODE) AND PRIORITY (1=HIGHEST PRIORITY).

Alert	Symbol	Alert context	Priority
Error alert	E	Feature	1
1 st level warning alert	W1	Feature	2
2 st level warning alert	W2	Working mode	3

Measurement. Data collection is performed by gathering Measures from Sensors and Machine control, that globally constitute the available Measurement tools. Each measure is tagged with a unique timestamp and represents the value of a specific Feature for a given Virtual Component, that operates in a specific Working mode.

State detection alerts. Alerts are raised when measures of one or more features, within a feature space that is being monitored, go beyond specific limits that are set according to the experience of domain experts, that are in charge of populating the model with these information for a given abstract virtual component. The same limits hold for all virtual components associated with the same abstract virtual component, that is, have been recognized as having the same behaviour for the same working conditions. In our model, we distinguish different kinds of alerts, namely warning and error ones, in turn organized at different levels, as summarised in Table I, where priorities among them are also provided (lower number means highest priority). Each feature has its own boundaries, that cannot be exceeded, whatever is the working mode. For these boundaries, we foresee a warning level, that raises a *first level warning alert* when it is violated, and is used to attract the focus on that feature. When the measures for a feature overcome the error boundaries, an *error alert* is launched on that feature. At lowest priority, features might present value boundaries also with regard to a specific working mode. If the measures exceed these boundaries, then a *second level warning alert* is launched. The rationale behind this choice is that, in a specific working mode, a feature usually assume values within a specific range. Values outside that range do not necessarily correspond to warning or error conditions for the feature, but might be an indicator that something is going wrong during the machine working. This corresponds to the second challenge highlighted in the introduction and is addressed here with multiple-level alert management. Since features are aggregated into feature spaces, alerts are associated to latter ones as well. In particular, an alert of type A_X (namely, error, first level warning, second level warning) is raised for a specific feature space if the same type of alert is raised for at least one of the features associated with the feature space. Furthermore, alerts are propagated taking into account the hierarchical aggregation of assets, as shown in Figure 2. We remark here that our data model does not constrain the application of any specific algorithm to establish if a feature goes beyond error or warning limits: this can be detected if a single value is out of limits, or after a given number of measures in a pre-defined time slot go beyond limits. This choice depends on the application domain and knowledge of experts who are in charge of populating the model.

IV. BIG DATA AS A SERVICE

Figure 3 depicts the functional architecture of the implemented service. The service provides methods to: (i) receive and collect data from machine fleet and store them according to the conceptual model proposed in Figure 2; (ii) push alert messages in case of error or warning conditions (see Table I); (iii) query the collected data, with focus on critical feature values within the scope of a specific working mode and a given feature space. The aim was at providing efficient data collection, ensuring at the same time prompt alerting and information storage and indexing apt to efficiently query the data infrastructure for critical feature values. This goal has been met as explained in the following.

After gathering measures from sensors and machine control, a pre-processing step is applied, where data are saved in a text file, containing on each row all the measures gathered in a given timestamp t_i , and cleaning techniques such as data filtering and quantization are applied to check the data quality and increase SNR (signal-to-noise ratio).

All information about the Virtual Component that is being monitored, the Working modes and associated Parameters, Features and Feature Spaces breakdown and warning limits are considered as configuration data, that are common for each Abstract Virtual Component. This information is stored within a MySQL relational DBMS, given the low variability and volume of this data. On the other hand, measures and their proper storage and management are the key point here and are stored within a NoSQL database (MongoDB): these data present a very simple structure, but are collected at high rate and their volume increases very quickly.

MongoDB stores data into documents, in turn organized into collections. This data organization is very flexible. Therefore, this flexibility can be exploited to further speed up querying over the database (see next section about performance experiments). In particular, in our approach, the structure of each document stored within MongoDB for each record of measures, collected in the same timestamp, is as follows:

Record ID	
Timestamp	
Working mode Params	Parameter 1
	Parameter 2
	...
	Parameter n
Feature Spaces	Feature 1.1 [E W1 W2]
	Feature 1.2 [E W1 W2]
	...
	Feature 1.m [E W1 W2]
	dots
	Feature k.1 [E W1 W2]
Feature Spaces	Feature k.2 [E W1 W2]
	...
	Feature k.h [E W1 W2]
	...

The basic idea is to process incoming cleaned data using MySQL configuration information, identify critical situations based on feature breakdown and warning limits (see Table I) and store information within MongoDB properly tagged with the type of alert (either error or warning) that has been recognized (E, W1, W2). If, as remarked in the previous section, alerts are raised after a given number of measures in a predefined time slot go beyond error or warning limits, tags are assigned, after raising alerts, to all measures that

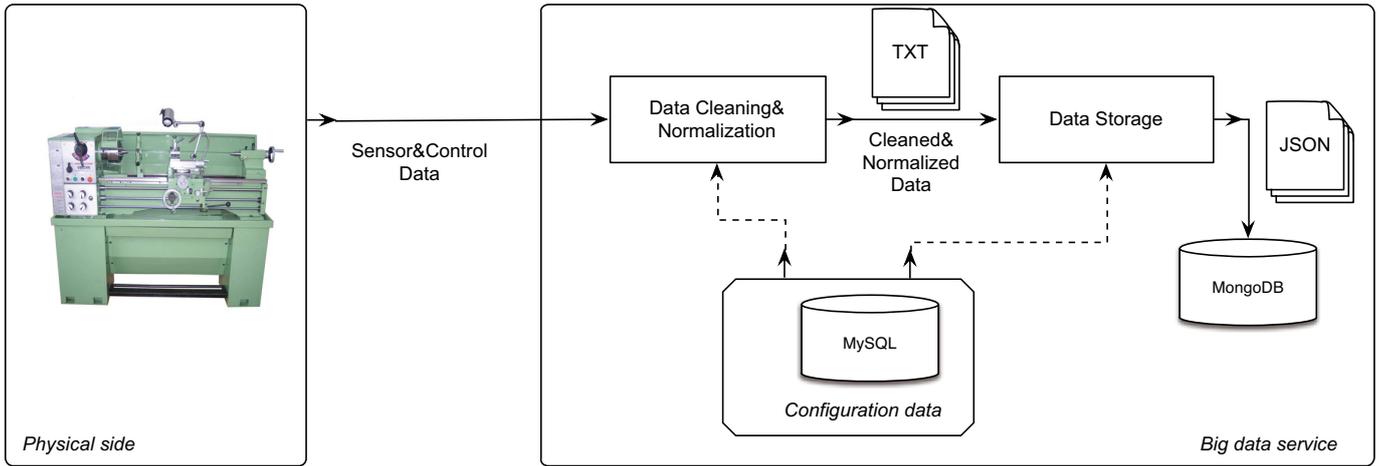


Fig. 3. Big Data as a Service for CPPS: functional architecture.

are out of limits. All these steps have been implemented within the *Data Storage* module (see Figure 3). This design choice slightly decreases performances in the acquisition phase compared to a straightforward plain MongoDB solution, where incoming records are directly stored as documents in a potentially infinite, although unique MongoDB collection. On the other hand, this contributes to speed up the querying phase, since all relevant information (including recognized critical situations) are already stored within the MongoDB storage space. Moreover, the organization of documents into MongoDB as implemented in our Big data service also takes into account relationships between *Feature measures*, *Feature Spaces* and *Working modes* (see Figure 2).

Experiments in Section V will demonstrate efficiency of our approach compared to a plain MongoDB solution and a storage and indexing solution that is completely based on MySQL.

A. Implementation issues

The Big data service is implemented in Java as a RESTful service [10], using XML as data exchange format. Implementation of the service is based on a distinct collection for each virtual component corresponding to a monitored physical asset or component. Therefore, collections are aggregated by abstract virtual component. MongoDB is an open source project developed in C++ and made available for all the well known platforms. The `mongo-java-driver-3.1.1` driver has been used to enable the interconnection between Java and MongoDB. JSON is the data format adopted to represent measures and save them within the MongoDB database. Listing 1 show a partial JSON representation for the case study in which the performances of Big data service have been tested.

In this example, three feature spaces are monitored for a spindle, namely axle hardening, shear stress and tool wear. Each feature space groups together a set of features: in the example, load and rpm of the spindle are used to monitor all the three feature spaces. Monitoring is performed also considering the working mode, that in this case is identified by the part program, the used tool and the mode (G0, i.e., quiescent state, in which the tool is mounted, or G1, in which

the spindle is working). A feature can also be tagged with the occurred alert type (namely, error E, first level warning W1, second level warning W2).

```

1 {
2   "_id" : ObjectId("565d9c4673738227a03a2244"),
3   "Timestamp" : "value",
4   "Parameters" : {
5     "PartProgram" : " value ",
6     "Tool" : " value ",
7     "Mode" : " value "
8   },
9   "FeatureSpaceList" : {
10    "AxleHardening" : {
11      "Load" : {"value":value,"alert":type},
12      "Rpm" : {"value":value,"alert":type},
13    },
14    "ShearStress" : {
15      "Current_X" : {"value":value,"alert":type},
16      "Current_Y" : {"value":value,"alert":type},
17      "Current_Z" : {"value":value,"alert":type},
18      "Speed_X" : {"value":value,"alert":type},
19      "Speed_Y" : {"value":value,"alert":type},
20      "Speed_Z" : {"value":value,"alert":type},
21      "Load" : {"value":value,"alert":type},
22      "Rpm" : {"value":value,"alert":type},
23    },
24    "ToolWear" : {
25      "Current_Z" : {"value":value,"alert":type},
26      "Speed_X" : {"value":value,"alert":type},
27      "Speed_Y" : {"value":value,"alert":type},
28      "Speed_Z" : {"value":value,"alert":type},
29      "Load" : {"value":value,"alert":type},
30      "Rpm" : {"value":value,"alert":type},
31    }
32  }
33 }

```

Listing 1. Example of JSON representation of measures

V. EXPERIMENTS

We tested the performances of the Big data service (BDaaS, Big Data as a Service) in terms of: (i) the number of records per second that can be loaded by the service (*acquisition rate*); (ii) the velocity in processing queries issued on the service (*querying rate*). The scalability of the approach has been checked by comparing the service implementation with other two versions: (a) an implementation based on a traditional relational DBMS (MySQL), where the whole conceptual schema shown in Figure 2 is engaged to design the logical schema

TABLE II. ACQUISITION TIME OF THE BIG DATA SERVICE (BDaaS), COMPARED AGAINST A PLAIN MONGODB IMPLEMENTATION AND A FULL MYSQL IMPLEMENTATION.

Num. of loaded records	Acquisition time in sec		
	BDaaS	Plain MongoDB	MySQL
2000	2,07 (0,0010)	1,62 (0,0008)	4,90 (0,0025)
4000	3,88 (0,0010)	3,45 (0,0009)	10,65 (0,0027)
8000	7,80 (0,0010)	6,30 (0,0008)	17,63 (0,0022)
12000	11,92 (0,0010)	9,44 (0,0008)	28,23 (0,0024)
16000	15,44 (0,0010)	13,35 (0,0008)	37,48 (0,0023)
20000	19,39 (0,0010)	16,95 (0,0008)	46,74 (0,0023)
30000	28,77 (0,0010)	23,48 (0,0008)	69,88 (0,0023)
40000	38,84 (0,0010)	32,63 (0,0008)	93,02 (0,0023)
60000	56,36 (0,0009)	48,68 (0,0008)	139,30 (0,0023)
80000	75,33 (0,0009)	62,36 (0,0008)	185,58 (0,0023)
100000	97,86 (0,0010)	73,13 (0,0007)	254,86 (0,0026)
150000	145,46 (0,0010)	120,90 (0,0008)	323,01 (0,0022)
200000	199,19 (0,0010)	153,41 (0,0008)	391,15 (0,0020)
400000	385,55 (0,0010)	332,64 (0,0008)	663,73 (0,0017)
800000	768,90 (0,00099)	608 (0,0008)	1360,44 (0,0017)
Avg acquisition time	0,0010	0,00086	0,00225

(i.e., tables) of the relational database; (b) an implementation of the service based on MongoDB, where all incoming records are simply stored within the same, huge MongoDB collection. The experiments have been carried out on a an Intel Core i7 platform, with 2.8 GHz CPU, 16GB RAM and Mac OS.

Table II shows acquisition time required by the three implementations to process incoming records. As expected, the MySQL-based solution is outperformed by the other two implementations. The plain MongoDB implementation slightly outperforms our implementation with reference to the acquisition time (since the loading procedure is simplified). Figure 4 presents a similar comparison in terms of acquisition rate (i.e., number of records per seconds that can be processed) and confirms the previous considerations, as well as the scalability of all the implementations when required to process an increasing number of records.

Concerning tests executed on querying rate (Figure 5), we firstly identified four kinds of relevant queries to be issued on the data infrastructure:

- Q1) a SELECT query, to retrieve all records stored in the data infrastructure;
- Q2) a SELECT query with projection, to retrieve all values for a subset of features, independently from the feature space and the working mode;
- Q3) a query to retrieve, given a Feature Space, all measures that exceed error or warning limits;
- Q4) a query to retrieve, given a Feature Space, all measures that exceed error or warning limits for a specific Working mode.

With reference to the conceptual model shown in Figure 2, query Q1 can be expressed as

```
Q1:: SELECT * FROM Measure;
```

Similarly, query Q2 can be expressed as:

```
Q2:: SELECT * FROM Measure, Feature WHERE
(Measure.FK_Feature = Feature.ID_Feature) AND
((Feature.name = F1) OR (Feature.name = F2) OR
...OR (Feature.name = Fn))
```

where Measure.FK_Feature is the foreign key constraint on Feature, F₁, F₂, ...F_n are the names of the features whose measures have to be retrieved.

Query Q3 can be expressed as:

```
Q3:: SELECT * FROM Measure, Feature, Feature_Space WHERE
(Feature_Space.FeatureSpace_ID = Feature.FK_FeatureSpace)
AND (Feature_Space.name = FSX) AND
(Measure.FK_Feature = Feature.ID_Feature) AND
((Measure.value < Feature.lowerBound_warning) OR
(Measure.value > Feature.upperBound_warning) OR
(Measure.value < Feature.lowerBound_error) OR
(Measure.value > Feature.upperBound_error))
```

where Feature.FK_FeatureSpace is the foreign key constraint on Feature_Space, FS_X is the name of the Feature Space target of this query. Finally, query Q4 can be expressed as:

```
Q4:: SELECT * FROM Measure, Feature, Feature_Space
Working_mode, Feature_mode_boundaries WHERE
(Working_mode.Mode_ID = MX) AND
(Measure.FK_Working_mode = Working_mode.Mode_ID) AND
(Feature_Space.name = FSX) AND
(Feature_Space.FeatureSpace_ID = Feature.FK_FeatureSpace)
AND (Measure.FK_Feature = Feature.ID_Feature) AND
(Feature_mode_boundaries.FK_Feature = Feature.ID_Feature)
AND (Feature_mode_boundaries.FK_Working_mode =
Working_mode.Mode_ID) AND
((Measure.value <
Feature_mode_boundaries.lowerBound_warning) OR
(Measure.value >
Feature_mode_boundaries.upperBound_warning))
```

where M_X is the ID of the Working mode target of this query, FS_X is the name of the Feature Space target of this query, properties named as FK_* represent foreign key constraints as in the previous query types.

Tests on querying rate have been performed by issuing ten different queries of each kind and computing the average time values. Figure 5 shows querying rate for the queries Q1-Q4 with respect to the number of records in the data infrastructure. As expected for queries Q1 and Q2, MySQL is outperformed by the other two solutions based on Big Data technologies. Moreover, plain MongoDB implementation and our approach are comparable. For queries Q3 and Q4, further controls are needed in the plain MongoDB implementation, also requiring to access configuration database to retrieve error and warning limits. In our approach, controls on such boundaries are already performed at acquisition time, and output of this processing is stored within MongoDB (see Listing 1). On the other hand, in the full MySQL solution, all configuration data are saved within the same MySQL database together with measures, thus a unique (although complex) query has to be issued for Q3 and Q4. This explains why for Q3 and Q4 full MySQL solution outperforms plain MongoDB implementation. Moreover, MongoDB documents are organized according to Feature Spaces only in our BDaaS implementation, therefore no interaction at all is needed with configuration database. Our approach presents better performances compared to plain MongoDB solution, thus demonstrating that the difficulty of mixing multiple technologies in BDaaS is compensated by

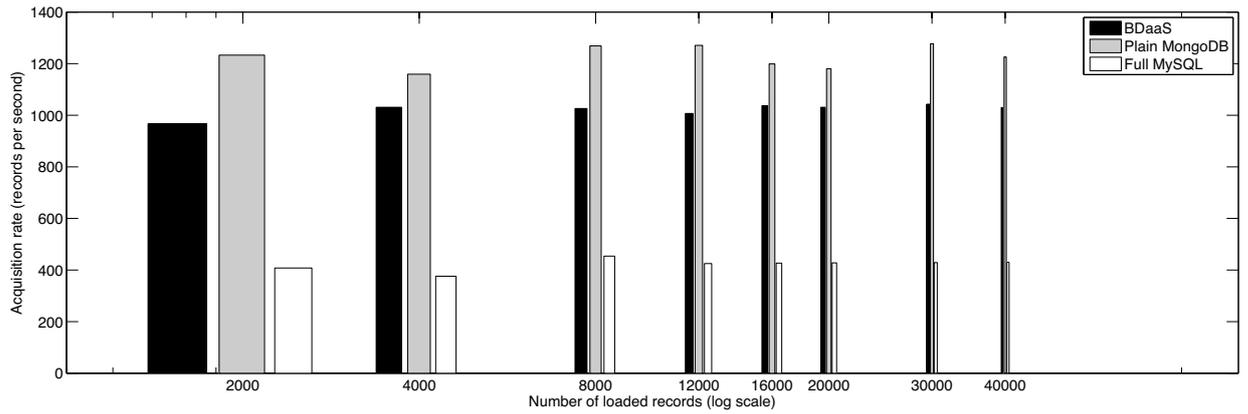


Fig. 4. Acquisition rate (number of collected records per second) for different implementations (BDaaS, plain MongoDB, full MySQL).

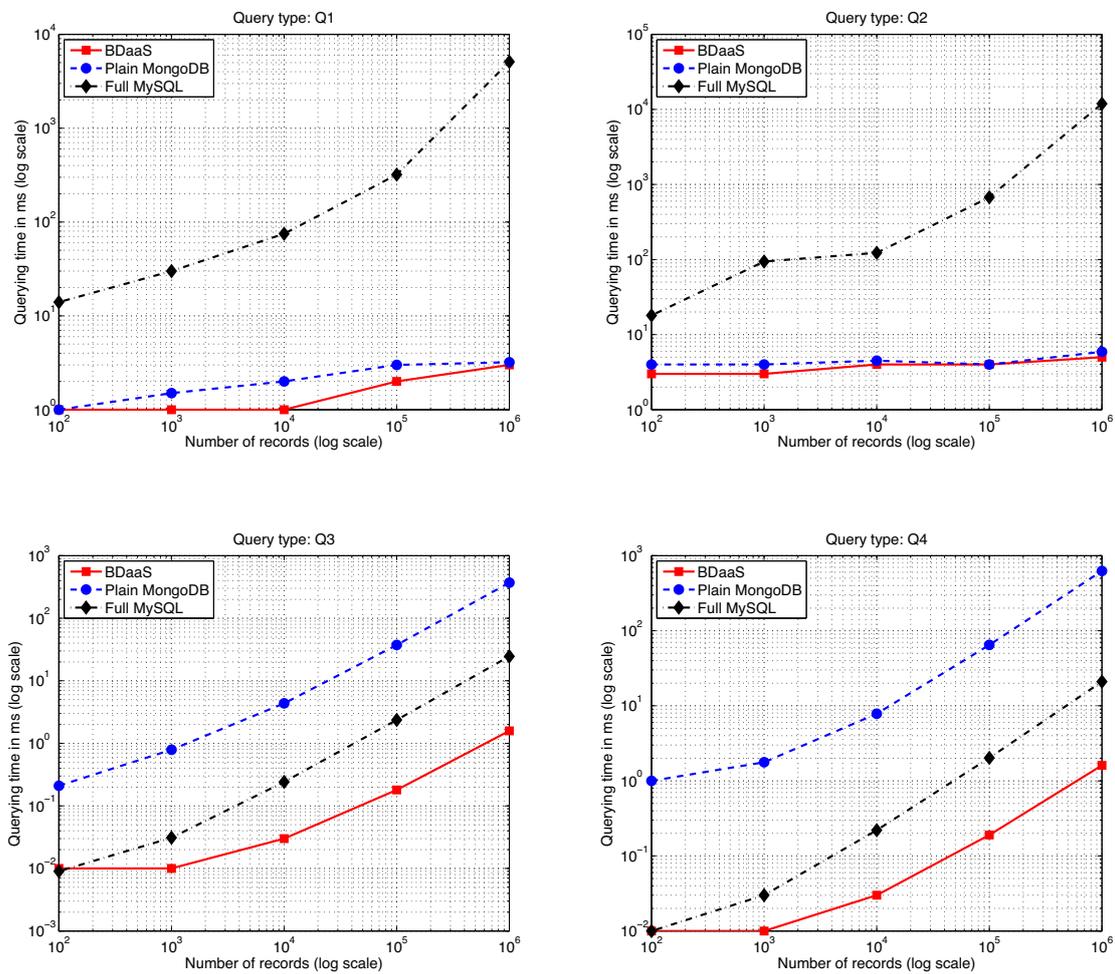


Fig. 5. Querying rate for different implementations (BDaaS, plain MongoDB, full MySQL) and different kind of queries.

performance improvements, taking the best from the different worlds.

VI. RELATED WORK

Traditional approaches on machine state detection are mainly focused on control-centric optimisation and intelligence. These approaches, and in general PHM (Prognostics and Health Management) systems, are simulation-oriented, that is, they rely on a separation between development and implementation, that is, PHM algorithms are based on lab generated and training data, that are different in quality and quantity from data collected from real machines and their surrounding environment (*in-field data*), that contain much more information. Moreover, they have been applied in specific fields, namely aircraft engines [12], wind turbine [4], electrical motors [1], machine tools [7] and so on. Our aim is not at solving a domain-specific problem, but at offering a flexible data infrastructure that is able to face challenges raised by the advent of Big Data technologies [11].

Research efforts have been devoted to the application of ontologies to the state detection problem [13]. Authors in [13] proposes an ontology, based on FMECA (Failure Mode, Effects and Criticality Analysis), that conceptualises the aggregation of parts of a wind turbine and is used as a reasoning base to propagate critical alerts occurring on sub-parts of the monitored system towards composite aggregations. The purpose of the proposed ontology is close to the hierarchical organization of assets in our model, but authors in [13] do not manage Big Data challenges and do not mention many of the aspects we highlighted in our model.

Available prognostic and state detection methods are mainly designed to support single machine monitoring and do not take advantage of considering similar machines as a fleet by gathering knowledge from similar, although different, instances of monitored assets. Novel methodologies are foreseeing the interaction between different surrounding systems, that pursue self-aware and self-learning machines by relying on observation of other assets operating in similar conditions [6]. We share with this approach the same vision, as also remarked by the same authors in [2], where Big Data era is promoted as carrier of big opportunities towards Manufacturing 4.0 full implementation. Nevertheless, in our paper we focused on the description of a data infrastructure acting as a fundamental engine for this revolution. Compared to recent efforts in storage service selection, also within modern cloud-based architectures [3], that support third-party storage service selection based on preferences of customers with respect to desired QoS, we provided here an optimised data acquisition and organization infrastructure rooted on Big Data technologies.

VII. CONCLUSIONS

In this paper, we describe a Data as a Service approach, that is able to manage data volume and velocity during the data collection phase, accumulating and summarizing measures from the machine fleet. Experiments on service performances demonstrate the efficiency of the proposed service, both in acquisition and in querying rate. The proposed Big Data as a Service approach constitutes a first step towards a full-fledged,

flexible and customizable architecture for CPPS-ization in the modern industry.

Next steps will concern the definition of a service-oriented infrastructure built on top of the proposed data infrastructure, including state detection and health assessment services. Within the design of such services, incremental and data stream clustering techniques will be studied, to identify working states of the monitored machines starting from the information stored within the data infrastructure described in this paper. This will enable a refinement of state detection algorithm, helping to identify rates of changes of monitored variables, in order to anticipate possible failures. We think that querying efficiency demonstrated by the Big Data service will be of paramount importance to this aim. Advanced automatic techniques to identify similarity across machines will be integrated in the state detection and health assessment services as well. Finally, security issues of cloud-based data collection for CPPS self-monitoring and representation of knowledge extracted from collected measures will be studied.

REFERENCES

- [1] B. Bagheri, H. Ahmadi, and R. Labbafi. Application of data mining and feature extraction on intelligent fault diagnosis by Artificial Neural Network and k-nearest neighbor. In *Proc. of XIX Int. Conference on Electrical Machines (ICEM)*, pages 1–7, 2010.
- [2] J. Lee, B. Bagheri and H.A. Kao. Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics. In *Proc. of Int. Conference on Industrial Informatics (INDIN)*, 2014.
- [3] C. Esposito, M. Ficco, F. Palmieri, and A. Castiglione. Smart Cloud Storage Service Selection Based on Fuzzy Logic, Theory of Evidence and Game Theory. *IEEE Transactions on Computers*, pages 1–14, 2015.
- [4] E. Lapira, D. Brisset, H.D. Ardakani, and D. Siegel. Wind turbine performance assessment using multi-regime modeling approach. *Renewable Energy*, 2012.
- [5] J. Lee, B. Bagheri, and H.A. Kao. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3:18–23, 2015.
- [6] J. Lee and H.A. Kao. Service innovation and smart analytics for industry 4.0 and big data environment. In *Proc. of 6th CIRP Conference on Industrial Product-Service Systems*, 2014.
- [7] L. Liao and J. Lee. Design of a reconfigurable prognostics platform for machine tools. *Expert systems with applications*, 37(1):240–252, 2010.
- [8] V. Martinez, M. Bastl, J. Kingston, and S. Evans. Challenges in transforming manufacturing organisations into product-service providers. *Journal of Manufacturing Technology Management*, 21(4):449–469, 2010.
- [9] L. Monostori. Cyber-physical production systems: Roots, expectations and R&D challenges. In *Proc. of the 47th CIRP Conference on Manufacturing Systems*, pages 9–13, 2014.
- [10] C. Pautasso, O. Zimmermann, and F. Leymann. RESTful Web services vs. “Big” Web services: making the right architectural decision. In *Proc. of the 17th Int. Conference on World Wide Web (WWW’08)*, pages 805–814, 2008.
- [11] A.B. Sharma, F. Ivancic, A. Niculescu-Mizil, H. Chen, and G. Jiang. Modeling and analytics for cyber-physical systems in the age of big data. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):74–77, 2014.
- [12] T. Wang, J. Yu, D. Siegel, and J. Lee. A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems. In *Proc. of Int. Conference on Prognostics and Health Management (PHM)*, pages 1–6, 2008.
- [13] A. Zhou, D. Yu, and W. Zhang. A research on intelligent fault diagnosis of wind turbines based on ontology and FMECA. *Advanced Engineering Informatics*, 29:115–125, 2015.

Atomic Instruction Translation towards a Multi-threaded QEMU

Alvise Rigo
Virtual Open Systems
Grenoble - France
a.rigo@virtualopensystems.com

Alexander Spyridakis
Virtual Open Systems
Grenoble - France
a.spyridakis@virtualopensystems.com

Daniel Raho
Virtual Open Systems
Grenoble - France
s.raho@virtualopensystems.com

KEYWORDS

TCG/QEMU; atomic instructions; system emulation; instructions emulation; parallel emulator; multi-threading

ABSTRACT

In the context of system emulation, the sophistication of the emulator usually grows with the complexity of the target system model. Particularly, emulating precisely a certain CPU architecture can introduce many challenges that have to be properly explored and somehow solved to reach an accurate emulation of the target system.

In this paper we present an implementation design of ARM atomic instructions for a multi-threaded version of QEMU (the *Quick EMUlator*), currently under development [1].

To prove the correctness and performance of such an implementation, some tests have been performed showcasing a high degree of accuracy and fidelity of the emulated instructions. While this paper does not cover all possible guest architectures that QEMU supports, the described new approach results in a reliable infrastructure that eventually can address all target architectures in QEMU.

I. INTRODUCTION

Multi-threading architectures have brought new challenges in the context of synchronization between parallel execution flows, requiring for specific multi-thread aware instructions.

Some of these instructions are called atomic, in that their execution is indivisible and uninterruptible. This means that any modifications to memory are always consistent, no matter how many CPUs can concurrently access it.

The main reason for such instructions is to simplify the code required to synchronize these flows, and of course to make it faster. Nowadays, almost every architecture has its own set of atomic instructions, which is often used to implement low-level synchronization routines.

These instruction sets, introduced complications for all system emulators, such as QEMU, that need to preserve the atomic nature of the instructions while emulating them. *TCG* (Tiny Code Generator), a software component parsing and translating the guest instructions to host instructions, has been originally implemented

with a single-threaded design in mind, although still capable of exposing multiple cores to the guest. In this simplified context, QEMU emulates multiple guest CPUs by executing them in a round robin fashion, making the guest actually slower than the uniprocessor variant.

Moving to a real multi-threading design (multiple guest CPUs executed concurrently in different threads) is a significant improvement for QEMU, allowing various emulation use cases in today's many-core systems. This design change comes with several challenges and among the most important ones is the proper and accurate translation of atomic instructions. With multi-threading in mind, atomicity is not granted by default and must be implemented carefully instead.

Challenges to be addressed

Emulation in itself brings two considerable challenges: accurate emulation for both the processor architecture, and the target machine model, including any attached devices/peripherals. The former challenge is of crucial importance, since in the context of QEMU it requires to:

- fetch the opcode instructions from guest memory
- decode and translate instructions to an intermediate code representation (*IR*), which is handled by the TCG frontend
- express the IR code in host machine instructions, which is done in the TCG backend.

The technique listed in the steps above is also called *Dynamic Translation* [Bellard 2005] and allows to create extremely portable full system emulators. For instance, adding support for a new guest/host architecture requires the implementation of a new TCG frontend and backend respectively, making the overall emulator design quite flexible and modular.

Every modern instruction set includes a number of atomic instructions which are extremely useful when implementing synchronization functions in a shared memory system. This kind of instructions are used to implement lock-free algorithms or, in general, programs where accesses to the shared data don't necessarily require a lock.

Examples of such instructions are the *Compare and Swap* (CAS) instructions (like the x86 `LOCK CMPXCHG` [2]) or the *LoadLink/StoreConditional* (*LL/SC*), introduced by Jensen, Hagensen, and Broughton [Jensen et al. 1988] as part of the *S-1 AAP* project.

The idea of the LL/SC instructions was to perform read-modify-write operations without requiring any bus to be locked, or in general, any CPU to be temporarily halted. This approach is superior compared to CAS instructions, because it solves the so called *ABA problem* when implementing non-blocking algorithms.

In a uniprocessor system all benefits brought by atomic instructions are irrelevant, as the implementation overhead of coherency for concurrent tasks is not needed. This is why QEMU was not concerned with such complications thus far, assuming that all guest instructions can be considered atomic by default. However, modifying QEMU towards real multi-threading (capable of emulating more than one CPUs at once) requires to revisit this simplified design, in favour of a more complex one, described in Section V of this paper.

II. ATOMIC INSTRUCTIONS

This section describes the semantics of the LL and SC instructions. Excluding minor differences, these instructions are considered semantically equal for all architectures.

LoadLink

This instruction reads the value from a shared memory location and stores the content into a register of the calling CPU. It also establishes a link and records the CPU with the accessed address (`xaddr`), to properly handle the subsequent SC operation. The LL marks the beginning of a *tentative exclusive memory region* [Jensen et al. 1988], that will be either confirmed or dismissed by the following SC instruction; each CPU defines its own EMR, only one at once is allowed. Depending on the architecture implementing the instruction, the size of the exclusive memory can be bigger than the size of the memory access (i.e. the size of the data read and then stored to a system register by the LL operation). For instance ARMv7 defines as IMPLEMENTATION DEFINED the *Exclusive Reservation Granule*, which is the size of the memory that will be monitored whenever a CPU issues a LL instruction.

StoreConditional

This instruction writes to the address `xaddr` only if it belongs to an exclusive memory region previously created by an LL. The SC is not always successful since another CPU can nullify the exclusive memory region by writing or reading to it. In general, the SC fails if a certain condition comes true. In its original definition, this condition has been defined specifically for the implementation of a particular processor architecture [Broughton et al. 1982], however, all recent architectures adopting these instructions implement a slightly different variant.

Invalidation of an exclusive memory region

Setting aside the actual implementation, this condition has to guarantee that the exclusive region initiated by the

LL has not been invalidated by any other CPU in the system. This includes any CPU capable of reading/writing to system memory, which can result in violating the initial assumption of an *exclusive* memory region.

From now on, the following notation will be used:

- $w_{P_i}(y, val)$ write of value val to address y made by process i
- $w_x(y, z, val)$ write access of size z bytes to address y made by processor x , the value written is val . When the value written is not relevant, the notation $w_x(y, z)$ will be used
- $r_x(y, z)$ read access of size z bytes made by processor x to address y . In some generic cases, for a read access to address x , the notation $load(x)$ will be used
- $ll_x(y, z)$ LoadLink instruction issued by processor x to address y , resulting in a read access of z bytes. In some generic cases, for a LoadLink to address y , the notation $loadLink(y)$ will be used
- $sc_x(y, z)$ StoreConditional instruction issued by processor x to address y , resulting in a write access of z bytes no matter what is the value written. In some generic cases, for a StoreConditional access to address y that writes the value val , the notation $storeCond(y, val)$ will be used
- $EMR_{x,y}$ exclusive memory region created by the system after CPU x performed $ll_x(y, z)$. $EMR_{x,y}$ persists until $sc_x(y, z)$ is performed.

III. THE ABA PROBLEM

The ABA problem usually occurs when a CAS-based non-blocking algorithm gives a false positive result [Dechev et al. 2010]. Listings 1 and 2 give an example of the same algorithm implemented using the two types of atomic instructions (CAS instruction and LL/SC instructions). The algorithm, called *updateValue*, updates the current value of a memory location at address `addr`. The algorithm makes use of the CPU's instructions CAS, LL and SC. In case of failure, SC returns 1.

Listing 1: `updateValue()` non-blocking algorithm with possible occurrence of the ABA problem

```

input: int addr
output: none
begin:
  do
    old ← load(addr)
    new ← compute_new_val()
    while CAS(addr, old, new) ≠ old
  end

```

If a process P_1 running $updateValue(addr)$ is interrupted after loading the value pointed by `addr` and before the execution of the CAS instruction, then the ABA problem will occur if a process P_2 , with $P_2 \neq P_1$, changes the value pointed by `addr` to old_1 , with $old_1 \neq old$, and eventually restore the value old (more concisely: $w_{P_2}(addr, old_1)$, $w_{P_2}(addr, old)$). In this scenario, the CAS instruction would succeed, without P_1 knowing that the value pointed by `addr` changed two times.

Listing 2: updateValue() non-blocking algorithm ABA problem resistant

```

input: int addr
output: none
begin:
  do
    old ← LL(addr)
    new ← compute_new_val()
    while SC(addr, new) = 1
  end

```

On the contrary, if the implementation provided by Listing 2 was used to handle the same scenario, there would not be any false positive since the SC would fail due to the first write $w_{P_2}(addr, old_1)$ made by process P_2 .

Suppose now that the code of Listing 2 is executed by a guest OS, and that TCG is used to translate the load-Link (LL) and storeCond (SC) instructions. Depending on the architecture of the host, TCG can rely on the same instructions of the host to directly map the guest instructions to host instructions. In ARM for instance, LL would be mapped to LDREX while SC to STREX.

However, for architectures such as x86 where similar instructions are not present, emulating correctly the LL/SC semantic is not obvious, and requires some additional effort. In other words, the emulation has to resolve the ABA problem using only CAS-like instructions. The simplistic emulation proposed by Listing 3 and 4 still suffers of the ABA problem, since in between the LL and SC, the value can be changed and then restored to *GLOBAL_old*.

Listing 3: LL emulation with CAS instruction

```

input: int *addr
output: int
begin:
  GLOBAL_old ← load(addr)
  return GLOBAL_old
end

```

Listing 4: SC emulation with CAS instruction

```

input: int *addr, int new
output: int
begin:
  written ← CAS(addr, GLOBAL_old, new)
  if new = written
    return 0
  else
    return 1
  end
end

```

In previous work [Dechev et al. 2010], it is proved already that it is possible to replicate correctly the LL/SC semantic without incurring the ABA problem, by the usage of additional variables which the non-blocking algorithm has to be aware of. For example [Gifford and Spector 1987], as representative of a common solution, makes use of version tags that pair the

actual data to be accessed: every access made by the CAS instruction would modify also the tag, guaranteeing **uniqueness** to every access. It is worth noting that the problem is solved only at an algorithm level, since spurious writes could still change only the value of the actual data, leaving the tag unaltered.

Given the nature of emulation, we can not make the guest aware of such an additional tag, and the extra care of emulating properly the LL/SC semantic has to be left entirely on the host. This makes the adoption of CAS-like instructions hardly feasible, especially if corner cases have to be handled. One such example is when 128bit wide LL/SC guest instructions have to be emulated relying on 32bit or 64bit cmpxchg host instructions, which would be the case for the emulation of ARMv8 LDXP/STXP instructions on x86_64. For reasons explained above, the emulation of atomic instructions can not be implemented in a straightforward way, imposing several challenges in which a major one is an actual solution to the ABA problem.

IV. QEMU INTERNALS

Before diving into details, some QEMU concepts are presented for completeness in the following paragraphs.

Instruction translation

In QEMU, the process of translating guest instructions to host instructions, is covered by TCG, which first translates guest instructions to an intermediate representation (defined by TCG instructions), and finally to a host representation. An example of this process is depicted in Figure 1, where TCG is translating ARM instructions to native (host) x86 instructions.

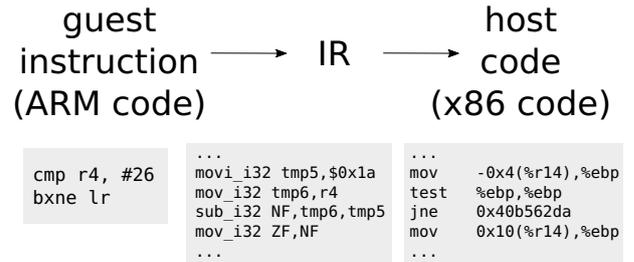


Fig. 1: Example of code translation for a TB

Host code is always contained in *Translation Blocks* (TBs), that although part of QEMU's address space, they are not normal functions but rather auto generated code of the QEMU process itself. Every TB contains a basic block of the guest code translated for the host machine. This generated code is not always enough to emulate all guest features (devices or instructions affecting the machine model state for instance); for this reason TCG allows to jump out of a TB to execute additional emulation functions. Figure 2 illustrates the TB structure, where the *prologue* and *epilogue* are the entry point and the exit point of the TB respectively.

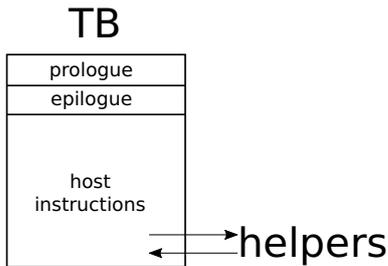


Fig. 2: Scheme of a TB

QEMU software MMU

QEMU offers two emulation modes, user and system. The first mode allows to run Linux user space applications and is not the topic of this paper. System mode instead is the most commonly used, but also the part of QEMU that receives most of the community effort. This mode of execution allows for full system emulation as the name suggests, relying on a software MMU (*softmmu*) that is able to solve the disparity between guest and host addresses. More specifically, the *softmmu* translates guest RAM (virtual) addresses to host pointers. The *softmmu*, in turn, relies on a TLB (Translation Lookaside Buffer) to cache the newly translated addresses (in the form of TLB entries). TLB entries are used to prevent guest table walks for addresses that are frequently accessed.

QEMU uses different bitmaps to track the different properties of each guest memory page. For example, one bitmap is used to distinguish memory mapped IO (MMIO) pages of the guest. These bitmaps are usually monitored when a new TLB entry is created, since the whole TLB is accessed directly from TCG code each time the guest wants to do a load or store operation. This happens mostly because reading and writing to memory requires translating the guest address into a host compatible, and the TLB serves exactly this purpose. While executing TB code, TLB entries are constantly checked to perform the translation; if the TLB entry does not exist or is meant for a different page, the execution exits the TB in order to generate the missing entry.

In other cases, e.g. MMIO pages, the corresponding TLB entries will always force the execution to leave the TB to execute, for instance, the emulation code of some device.

According to QEMU terminology, we will call *slow-path* the execution that exits from the current TB, either due to a missing TLB entry or for a TLB entry that requires an exit (as per MMIO entries). In all the other cases, the execution will follow the so called *fast-path*.

V. TOWARDS MULTI-THREADING

Extending QEMU’s infrastructure to properly translate atomic instructions in a real multi-threaded implementation, can result in two notable design options: One, where the translation of atomic instructions is fully in charge of the guest frontend, eventually using helpers to implement

additional steps that are not directly covered by TCG-generated code. This option would result in modifying all QEMU supported guest architectures in a consistent way, thus making the transition to multi-threading even more demanding. The other solution, proposed in this paper, aims at providing a unified code infrastructure for all guest frontends, providing an accurate implementation that abides more closely to the semantics dictated by the architecture specification.

LL/SC helpers

Two helper functions consist the core of atomic instruction translation. *LoadLink* and *StoreConditional*, which are functionally equal to the analogous instructions described in Section II. The two helpers have been designed to have an one to one mapping to the corresponding atomic instructions present in architectures such as ARM, that adopted the LL/SC paradigm. In other cases, for example x86 `LOCK CMPXCHG`, the instruction can be easily achieved by means of LL/SC instructions as presented in [Maged 2004] and [Anderson and Moir 1995]. Even if the work presented in this paper is focused on ARM `ldrex` and `strex`, it nevertheless provides the means to translate atomic instructions for all other architectures that are supported by QEMU.

These two helpers rely heavily on QEMU’s *softmmu*. Since being extremely deep-rooted in QEMU’s MMU emulation layer, it is the perfect tool to address all the features that the atomic instructions require. From now on, we will refer to the notation that has been introduced in Subsection II, where $ll_x(y, z)$ and $sc_x(y, z)$ map respectively to `ldrex` and `strex`, $w_x(y, z)$ and $r_x(y, z)$ map to normal load and store instructions (namely `ldr` and `str`).

The translation of $ll_x(y, z)$ has to define the correlated $EMR_{x,y}$, in such a way that the link established by the first instruction is honoured by all CPUs in the system. In practice, the system has to react properly to all write accesses made to $EMR_{x,y}$ invalidating the link. Figure 3a depicts what would happen if all the fast-path accesses are not taken into account.

To protect against this, a new bitmap, called *exclusive*, has been added to QEMU’s *softmmu*, which flags all the pages containing an active EMR. Whenever QEMU generates a TLB entry for a guest page, the corresponding *exclusive* bit in the bitmap is checked. A set bit on a page will make all the TLB entries created for that page to be such that, if evaluated by the guest code for an address translation (something that occurs for every $r_x(y, z)$ and $w_x(y, z)$), they will always force the execution to exit from the TB. This will be used as a hook to trap from a TB to a QEMU function that evaluates if the access conflicts an existing EMR. If such a conflict occurred, the subsequent SC instruction will fail. In any case, the normal write access that triggered the failure will always succeed. Accesses to pages containing EMRs are depicted in Figure 3a.

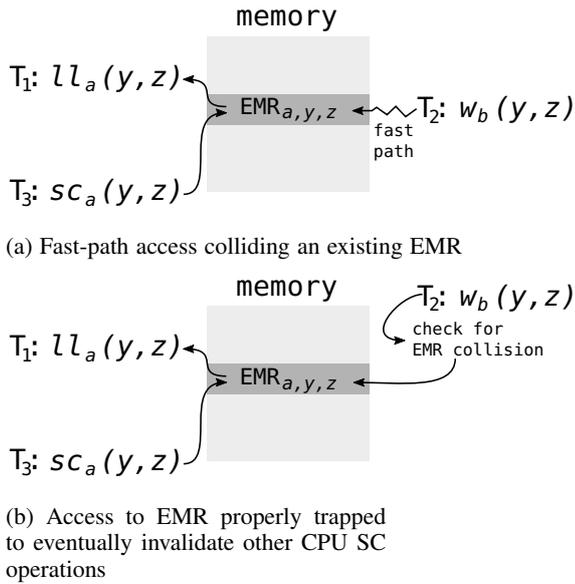


Fig. 3: In both (a) and (b) $T_1 < T_2 < T_3$ is applicable. In (a) guest CPU 2 is writing to the EMR through the fast-path, the following $sc_a(y, z)$ will anyhow succeed. In (b) access $w_b(y, z)$ makes CPU 2 evaluate a TLB entry that will require to leave the fast-path

The definition of the EMR is done by storing its range inside the guest CPU state variable: this is perfectly in line with the impossibility to have nested LL/SC operations like $ll_x(y_1, z), ll_x(y_2, z), sc_x(y_1, z), sc_x(y_2, z)$.

VI. IMPLEMENTATION DETAILS

In this section all complications regarding multi-threading execution will be described, together with the design choices of the proposed implementation. One major problem when dealing with multi-threaded programs is the occurrence of race conditions. In the context of this work, a race condition can be associated to an inconsistency of the whole machine state, which is in charge of translating atomic instructions. The direct negative result of such a state is the failure of a SC operation that should have succeeded, or even worse, the success of a SC operation that had to fail. In the following sections, all critical points that result in race conditions are explored, where the implemented approach is also documented.

Updates of the exclusive bitmap can lead to inconsistencies due to the out-of-order execution of load/store operations as seen, for instance, on ARM architectures [3]. For this reason all accessors to such a bitmap are atomic, an outcome that is possible by means of host atomic instructions. It is important to note, that this can be possible only in the case where bitmap accessors are QEMU functions and not implemented through TCG generated code.

Setting a bit in the exclusive bitmap to enforce slow path execution for the TLB entries overlapping an EMR, can result in another problem. In fact, other guest CPUs, different from the one issuing the LL, could have already

generated TLB entries for the same page, forcing the execution to follow the fast-path (as what happens in Figure 3a). To avoid this dangerous behaviour, TLB entries of these CPUs will be flushed, forcing them to recreate the TLB entry that covers the page in the EMR. This flush request will also prevent race conditions, that are related to the delayed new state propagation of the exclusive bit.

Lastly for this implementation, the evaluations and updates of the EMRs have been safeguarded using a mutex. This is mandatory because updating this structure is not possible with a single atomic instruction. Another related aspect that requires additional caution, relates to the actual memory accesses made by the LL and SC instructions. More specifically, the results on memory brought by these instructions has also to be done jointly with the update of the EMR values. The Listings 5, 6 and 7 represent respectively the LL, SC and normal store access. In these examples, the critical region is delimited by two calls LOCK and UNLOCK.

For instance, consider the Listing 5, which only works as long as the normal load is done inside the critical section, otherwise the loaded value can be potentially updated by another CPU, which might or might not be inside the critical region. For the same reason, the SC operation (Listing 6) has also to rely on the same critical region to be consistent with the rest of the atomic instruction emulation. Without entering the critical region, it can potentially declare the operation as successful (returning 0), but performing the store after another CPU modified the value. Similarly, the store operation (Listing 7) enters the critical region to check for a possible conflict in EMR, but also to perform the regular access.

Listing 5: LoadLink pseudo code, `load()` denotes a plain load from memory of size z

```

input: int y, int z, int x
output: int
begin:
    LOCK()
    CPU[x].EMR ← [y, y + z]
    ret ← load(y, z)
    UNLOCK()
return ret
end

```

Listing 6: StoreCond pseudo code, `store()` denotes a plain store to memory of size z

```

input: int y, int z, int x,
        int val
output: int
begin:
    LOCK()
    if CPU[x].EMR = [y, y + z]
        store(y, z, val)
        ret ← 0
    else

```

```

    ret ← 1
end
UNLOCK()
return ret
end

```

Listing 7: Plain write access trapped by the slow-path

```

input: int y, int z, int val
begin:
    LOCK()
    for each CPU
        if CPU.EMR overlaps [y, y + z]
            CPU.EMR ← NULL
        end
    store(y, z, val)
    UNLOCK()
end

```

VII. EXPERIMENTAL RESULTS

The implementation has been evaluated under two main points of view: correctness of the emulation and performance. Both of them are relevant aspects that have to be properly considered: from one side, the translation of atomic instructions needs to behave in accordance with the architecture instruction set specification. From the other side, the performance evaluation has also to be taken into consideration, as too much overhead would slow down the guest execution considerably. Atomic instructions in fact, are mostly used for synchronization routines, e.g. the Linux implementation of the spin lock functions, and are designed to keep as short as possible the lifespan of the EMR in order to minimize the number of conflicts.

[Wang et al. 2011] and [Ding et al. 2011] could have been two candidates for comparison against the presented work, however, only the upstream version of QEMU (whose source code is available at [4]) will be taken into account. In fact, QEMU has been significantly evolving over the last years and as such, results from these previous attempts would be considered deprecated and outdated. In addition, as it has been described in Chapter V, the presented implementation sets the objective of providing a common infrastructure for atomic instruction translation for all supported architectures, relying as much as possible on the current QEMU code base, as well as its features and components. While the proposed implementation results in a slightly greater overhead than a guest specific implementation, at the same time it is also more beneficial to the QEMU community by offering a unified implementation and a faster upstreaming process.

ABA problem occurrence test

With the purpose of verifying that the ABA problem could actually occur in the context of emulation (as explained in Section III), a specific test has been implemented. The test is used as a proof that LL/SC instructions can not be translated by relying only on a

straight application of the host’s CAS instruction. The C code of the test is presented in Listing 8, each thread is pinned to one ARMv7 guest CPU. The value of `ADDEND` has been chosen explicitly to make the identification of the ABA occurrence easy to verify on the host emulation side, not in the guest. In fact, in case the problem occurs, the STREXD instruction will fail as it should, *but it would not fail in the case* `ADDEND` *was 1* (in such a case, the occurrence would not be noticed by the guest). Listing 9 shows the naive implementation of the STREXD instruction used in the multi-threaded QEMU code; the method `lsb32` returns the 32 least significant bits of the argument.

Listing 8: ABA problem triggering test

```

/* shared counter */
uint64_t global_cnt

/* thread 1 */
for (i = 0; i < LOOP_SIZE; i++) {
    __sync_fetch_and_add(&global_cnt, 2);
}

/* thread 2 */
#define ADDEND ((1 << 32) | 1)
for (i = 0; i < LOOP_SIZE; i++) {
    __sync_fetch_and_add(&global_cnt,
                        ADDEND);
    __sync_fetch_and_sub(&global_cnt, 1);
}

```

Listing 9: ABA problem: STREXD emulation code

```

input: int *addr, int new
output: int
begin:
    written ← CAS(addr, GLOBAL_old, new)
    if new = written
        return 0
    else
        if lsb32(written) = lsb32(new)
            aba_errors++
        end
        return 1
    end
end

```

The cycle we need to identify is (here `ll` does not report the access size, the second argument of `sc` is the value written): $ll_1(addr) \rightarrow ll_2(addr) \rightarrow sc_2(addr, val_1 + ADDEND) \rightarrow ll_2(addr) \rightarrow sc_2(addr, val_2 - 1) \rightarrow sc_1(addr, val_1)$, where val_1 is the value retrieved by both $ll_1(addr)$ and the first $ll_2(addr)$, $val_2 = val_1 + ADDEND$ and $addr$ is the address of the global counter. More specifically, the first and last instructions implement the `__sync_fetch_and_add` of thread 1, the inner instructions instead implements `__sync_fetch_and_add` and `__sync_fetch_and_sub` of thread 2. When such a cycle happens, the 32 most significant bits are incremented by $(1 \ll 32)$, while the least significant half returns to the original value (e.g.: `0x00000001 00000001` \rightarrow `0x00000010 00000010` \rightarrow `0x00000010 00000001`): when this particular sequence of values takes place, the code registers the event incrementing the ABA errors counter.

<i>LOOP_SIZE</i>	occurrence (%)
10x10 ⁶	0.0008
8x10 ⁶	0.0008
6x10 ⁶	0.0007
2x10 ⁶	0.0005
1x10 ⁶	0.0004

Table 1: Occurrence ratio of ABA problem (referring the test code of Listing 8)

Table 1 reports the percentage of errors according to the value of `LOOP_SIZE`, verifying the actual risk derived from an improper translation of LL/SC instructions. It is worth to note that thread 2 could have used plain load and store accesses; this would have probably made the occurrence of the ABA flaw even more consistent, though greatly complicating the algorithm to detect its occurrence.

Benchmarking tests

A benchmark bare metal application has been developed to evaluate the correctness and performance of the proposed solution, which is compared against the current version of QEMU. The application is also a stress test, as it puts the whole machinery under heavy load, implementing a scenario characterized by significant contention of a shared resource. The source code has been kept at a bare minimum and designed without any operating system dependencies.

Listing 10: Code used to benchmark the emulation overhead, executed for every core of the guest.

```

for i ← 0 to LOOP_SIZE
  LOCK()

  if global_a = cpu_index % 2
    global_a ← 1
    global_b ← 0
  else
    global_a ← 0
    global_b ← 1
  end

  if global_a = global_b
    errors ← errors + 1
  end

  UNLOCK()
end

```

Listing 11: Simplified ARM assembly code of the `LOCK()` function. The register `r2` contains the address of the shared lock.

```

lock:
mov r1, #1
repeat:
ldrex r3, [r2]
strex r0, r1, [r2]
cmp r0, #0
bne repeat
cmp r3, #0
bne repeat

```

In Listing 10 the pseudo code of the stress test is reported, the variable `global_a` and `global_b` are shared among all the processors, while the functions `LOCK` and `UNLOCK` are used to define the critical region, providing means to hold/release a basic lock.

The purpose of the `if` statement is to swap the two values according to the index (identifier) of the CPU that has the lock. The aim of the test is to verify the correctness of the execution: in case the critical region was not respected, then the assignments to the global variables would overlap with the likely outcome of the two variables being set to the same value.

The implementation of `LOCK` can be found in Listing 11; as far as `UNLOCK` is concerned, its implementation is not reported as it resolves in a normal store to memory. Given the particular design of the test, the multi-threading feature of QEMU doesn't offer any advantage since the single iteration of the test case is spent almost entirely inside the critical region.

The results of the stress tests are reported in Figure 4a and 4b where `LOOP_SIZE` is respectively one million and ten million. The number of guest CPUs (reported in the x-axis) was always lower or equal to the number of host CPUs.

The host machine used for the tests is an Intel Core i7-4710MQ clocked at 2.5GHz while the machine model used in QEMU is `ARM virt`.

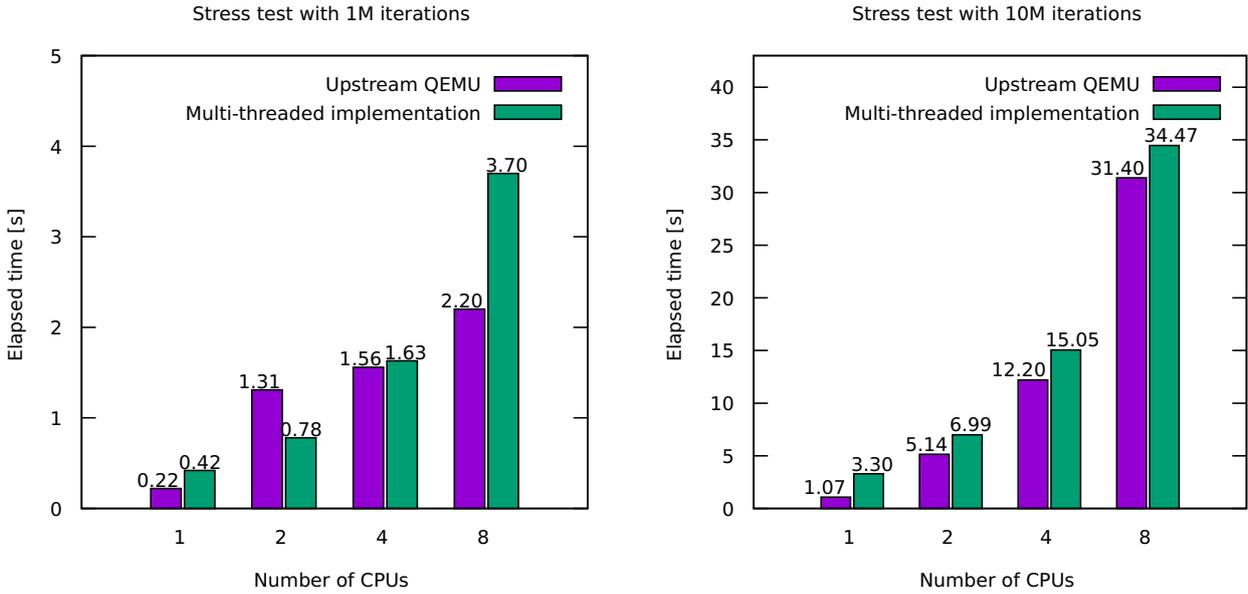
With the aim of testing the code in a real use case scenario, the boot time of the Linux kernel has also been measured. The results are reported in Figure 5.

Results analysis

By implementing a test designed to showcase the ABA problem, the risk of false positives in the field of atomic instruction emulation is confirmed. The same test could not have been possible with the implementation object of this work, since its overall design guarantees the eviction of the ABA problem from the start. However, the most complicated implementation of atomic instruction translation based on a multi-threaded QEMU is slightly slower than the upstream counterpart. In upstream QEMU there is no real CPU concurrency and all the iterations are serialized inherently by QEMU, through its round-robin scheduling of guest CPUs. Moreover, the emulation of `ldrex` and `strex` is done entirely inside a TB, which is a huge performance advantage since it does not require any branches to C code. As we can see in 4a, the relatively low number of iterations pronounces the overhead by the proposed implementation significantly. However, Figure 4b shows how the gap between the two implementations is lower with an increased number of iterations. With ten million iterations the overhead introduced by the new atomic translation is less than 9%. The stress test has also proved the correctness of the emulation, since the error counter never left the initial value of 0.

In the case of Linux kernel boot-up times, the overhead with one guest CPU increases the boot time by 13.5%. However, as soon as the benefits of multi-threaded

Fig. 4: Stress test result



(a) Stress test result with one million iterations

(b) Stress test result with ten million iterations

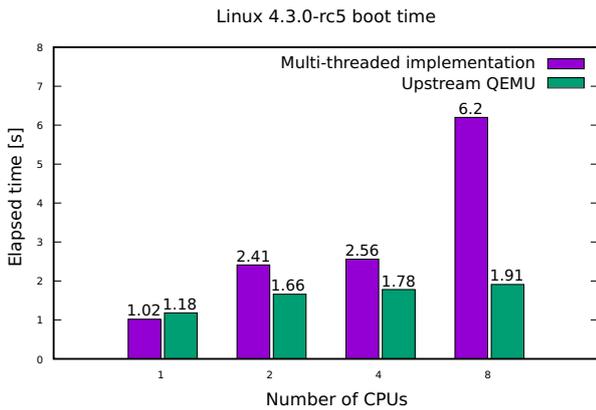


Fig. 5: Boot time of the Linux kernel

QEMU come into play, upstream QEMU is outperformed as we can see from Figure 5.

VIII. RELATED WORK

There have been already some attempts to make QEMU multi-threading in the past, all of them had to address the problem of atomic instruction translation. PQEMU [Ding et al. 2011] implemented the translation of ARM’s LDREX and STREX instructions enclosing their actual memory access in a mutex. This solution, although capable of serializing all the overlapping atomic accesses, is unable to emulate completely the LL/SC semantic since it does not solve the ABA problem. In fact, in the sequence $ll_a(y, z, val_1) \rightarrow w_b(y, z, val_2) \rightarrow w_b(y, z, val_1) \rightarrow sc_a(y, z, val_3)$ the SC does not fail. COREMU [Wang et al. 2011] also does not address at all the ABA problem, without even handling concurrent LDREX or STREX instruction.

Other related works like [Hong et al. 2012], [Chipounov and Candea 2010], [Scheller 2008] and [Brad et al. 2012] combined (or entirely replaced) the current dynamic binary translator of QEMU with LLVM, offloading to an additional thread the compilation and optimization of the generated code. These works, although proposing indeed a multi-threaded version of QEMU, keep the emulation of multiple guest cores in one single thread, as in the vanilla version of QEMU.

Broadening the scope, another work that relates to QEMU, since it adopts dynamic binary translation, is ARCSim [Almer et al. 2011] that translates the ARCOMPACT atomic exchange wrapping its emulation with spin locks. While this solution can work to emulate CAS-like instructions, it would not work with LL/SC instructions. Similar conclusions can be drawn for [Almer et al. 2012] that instead maps the ARCOMPACT atomic instruction directly to x86 compare and exchange. This is perfectly fine for the scope of the ARCOMPACT ISA, but would not work for LL/SC instructions emulation due to the reasons presented in Listing 3 and 4.

These related works can certainly give some hints on how to translate specific guest atomic instructions, but do not help in building an infrastructure that could address different types of guest instructions emulated by different host architectures.

IX. CONCLUSION AND FUTURE WORKS

This paper described the key points for a new unified approach towards atomic instruction translation in the QEMU emulator, focusing on the set of problems introduced by parallel execution of guest cores, like the ABA problem.

One of the most important aspect discussed, relates to the use of the LL/SC semantic to implement a set of helper functions. These helpers can be used to translate ARM's *ldrex* or *strex* directly, but also additional ones e.g. CAS instructions.

As expected, the overhead of the emulated instructions turned to be higher than the non multi-threaded counterpart, but still under 15% in a real use case scenario. Important to note is that the proposed implementation did not produce any errors and resulted in a flawless boot of Linux (which makes considerable use of atomic instructions). While there is a cost in performance, the new infrastructure provides means for a real parallel execution of SMP guest environments, as we can see in Figure 5.

The source code for the proposed implementation, which is currently at its seventh iteration, is available at [5] git repository. The patch series will be updated according to the feedback of the community that already helped to advance the work to this stage. In the future releases, the support of other architectures will be added to extend the multi-threaded execution for further QEMU supported targets.

X. ACKNOWLEDGMENT

The authors of this paper would like to thank Huawei Technologies Duesseldorf GmbH that made possible the realization of this work.

REFERENCES

- [Almer et al. 2011] O. Almer, I. Bohm, T. von Koch, B. Franke, S. Kyle, V. Seeker, C. Thompson, and N. Topham, *Scalable multi-core simulation using parallel dynamic binary translation*, in Embedded Computer Systems (SAMOS), 2011 International Conference on, July 2011, pp. 190-199
- [Almer et al. 2012] Oscar Almer, Igor Bhm, Tobias Edler von Koch, Bjrn Franke, Stephen Kyle, Volker Seeker, Christopher Thompson, Nigel Topham, *A Parallel Dynamic Binary Translator for Efficient Multi-Core Simulation*, International Journal of Parallel Programming, 2012, Volume 41, Number 2, pp. 212-235
- [Anderson and Moir 1995] James H. Anderson and Mark Moir. 1995. *Universal constructions for multi-object operations*, fourteenth annual ACM symposium on Principles of distributed computing (PODC '95). ACM, New York, NY, USA, 184-19
- [Bellard 2005] Bellard, Fabrice. *QEMU, a Fast and Portable Dynamic Translator*, In USENIX Annual Technical Conference, FREENIX Track, pp. 41-46. 2005.
- [Brad et al. 2012] Brad Alexander, Sean Donnellan, Andrew Jeffries, Travis Olds, and Nicholas Sizer. 2012. *Boosting instruction set simulator performance with parallel block optimisation and replacement* In Proceedings of the Thirty-fifth Australasian Computer Science Conference - Volume 122 (ACSC '12), Mark Reynolds and Bruce Thomas (Eds.), Vol. 122. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 11-20.
- [Broughton et al. 1982] J. M. Broughton, P. M. Farmwald, T. M. McWilliams, *S-1 Multiprocessor System*, SPIE Technical Symposium East, Arlington, Virginia, May 3-7, 1982
- [Chipounov and Candea 2010] Chipounov, Vitaly, and George Candea, *Dynamically Translating x86 to LLVM using QEMU* No. EPFL-REPORT-149975. 2010.
- [Dechev et al. 2010] Damian Dechev, Peter Pirkelbauer, Bjarne Stroustrup, *Understanding and Effectively Preventing the ABA Problem in Descriptor-based Lock-Free Designs*, 16th IEEE International Symposium (ISORC 2013), pp. 185-192, 2010
- [Ding et al. 2011] Jiun-Hung Ding, Po-Chun Chang, Wei-Chung Hsu, and Yeh-Ching Chung. 2011. *PQEMU: A Parallel System Emulator Based on QEMU*, ICPADS '11, Washington, DC, USA, 276-283
- [Gifford and Spector 1987] David Gifford and Alfred Spector. *Case study: IBM's system/360-370 architecture.*, Commun. ACM 30, 4 (April 1987), 291-307, 1987
- [Hong et al. 2012] Hong, Ding-Yong, Chun-Chen Hsu, Pen-Chung Yew, Jan-Jan Wu, Wei-Chung Hsu, Pangfeng Liu, Chien-Min Wang, and Yeh-Ching Chung, *HQEMU: a multi-threaded and re-targetable dynamic binary translator on multicores* In Proceedings of the Tenth International Symposium on Code Generation and Optimization, pp. 104-113. ACM, 2012.
- [Jensen et al. 1988] E. H. Jensen, G. W. Hagensen, J. M. Broughton, *A New Approach to Exclusive Data Access in Shared Memory*, S-1 Project. 15th Annual International Symposium on Computer Architecture, 1988
- [Maged 2004] Maged M. Michael, *ABA Prevention Using Single-Word Instructions*, IBM Research Division, RC23089, Tech. Rep., January 2004
- [Scheller 2008] T. Scheller. *LLVM-QEMU*, Google Summer of Code Project, 2008.
- [Wang et al. 2011] Zhaoguo Wang, Ran Liu, Yufei Chen, Xi Wu, Haibo Chen Weihua Zhang and Binyu Zang. *COREMU: a Scalable and Portable Parallel Full-system Emulator*, ACM SIGPLAN PPoPP 2011. San Antonio, USA, February, 2011
- [1] *QEMU's wiki web page*, <http://wiki.qemu.org/Features/tcg-multithread>
- [2] *Intel 64 and IA-32 Architectures Software Developers Manual*, <http://www.intel.fr/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-instruction-set-reference-manual-325383.pdf>, September 2015
- [3] ARM white paper *Memory access ordering - an introduction*, <http://community.arm.com/groups/processors/blog/2011/03/22/memory-access-ordering-an-introduction>
- [4] *QEMU git repository* [git://git.qemu-project.org/qemu.git](http://git.qemu-project.org/qemu.git)
- [5] Source code GIT repository: https://git.virtualopensystems.com/qemu_tcg/mmtcg_qemu/tree/slowpath-for-atomic-v5

Towards Secure Non-Deterministic Meta-Scheduling for Clouds

Agnieszka Jakóbiak, Daniel Grzonka, Joanna Kołodziej
Institute of Computer Science
Faculty of Physics, Mathematics and Computer Science
Cracow University of Technology

Horacio González-Vélez
Cloud Competency Center
School of Computing
National College of Ireland

KEYWORDS

Scheduling; Genetic algorithms; Cloud security; Task farm; Algorithmic skeletons; Agents; Parallel processing; Cloud computing

ABSTRACT

Task scheduling in large-scale distributed High Performance Computing (HPC) systems environments remains challenging research and engineering problem. There is a need of development of novel advanced scheduling techniques in order to optimise the resource utilisation. In this work, we develop the Agent Supported Non-Deterministic Meta Scheduler for cloud environments. This scheduling model is a simple combination of intelligent agent-based monitoring model for cloud system and security-aware cloud scheduler. In our model, scheduling, monitoring and reporting are provided in non-deterministic time intervals. An empirical case study using a FastFlow task farm was presented. It has demonstrates the effectiveness of the proposed solution.

I. INTRODUCTION

Cloud computing became a key paradigm for massive data processing and large scale computing. Scalability, flexibility, virtualisation, and geographical distribution of resources and users in computational clouds are the main features of such systems. The efficient management of available resources, in form of task scheduling, is the key point in every cloud system. Optimal task scheduling of available resources becomes even more challenging when considering security constraints.

In computational clouds, pervasive virtualisation of distributed resources (e.g. networks, servers, storage, applications,...) supports the distributed access to the cloud services. The cloud service can be therefore defined by a set of personalised functionalities provided by virtualised resources [16]. Canonical cloud services class include the following components:

- **Infrastructure as a Service (IaaS)** which encompasses hardware, network, storage, and operating system, typically packaged as a Virtual Machine (VM) or instance;
- **Platform as a Service (PaaS)** which provides an integrated development environment with control over the

deployed applications and environment configurations; and

- **Software as a Service (SaaS)** where final users employ software via application programming interfaces and browsers.

In order to expose cloud services as a utility [5], cloud providers must deploy resource pooling, rapid elasticity, on-demand self-service, measured service, and broad network access. It is therefore crucial to efficiently administrate and provision computational resources to meet user needs.

Traditional cloud scheduling strategies usually focus on the Makespan reduction. Additional criteria such as energy consumption, security-awareness, load balancing, reliability, intercommunication time, heterogeneous elements (resources and/or tasks), proper completion time, and service level agreements significantly increase complexity of scheduling. Therefore, cloud scheduling can be defined as a family of NP-complete optimisation problems. Depending on the users needs, the complexity of the scheduling problem may be determined by the number of objectives to be optimized, the type of the environment, and/or the processing mode and tasks interrelations [14].

Many concepts of the multi-criteria cloud schedulers are published in the literature. The List Scheduling [3] algorithm is an example of the simple greedy algorithm minimising the makespan: whenever a VM becomes available it process any unprocessed job. The Least-Cost-Last algorithm [11] allows to minimise given objective: after finding the job that has the latest due date, it recursively schedules all the remaining jobs to minimise a given objective. Randomized rounding and linear and dynamic programming algorithms are used to solve the multi-objective problem [4]. The most popular rounding algorithm is the Round Robin Algorithm [7], and its modified versions: Weighted Round-Robin [9], Fair Round-Robin [22] and Adaptive Round Robin [17].

While soft methods based on evolutionary computing (e.g. genetic algorithms) have been documented [14], Independent Batch Scheduling (IBS) offers a substantial improvement [15]. IBS groups tasks into batches which can be executed independently. The paper considers IBS problem in the context of increasing security and timing of system monitoring.

Cloud environments may also be overexploited by unauthorized individuals and organisations. Without the proper control

one would be able to upload his own task. Then the cloud system will work on his benefit and the dishonest will not pay for computational time. To control the privileges users have different roles with distinct access to resources. Schedulers and service providers must also take into account the privileges, internal security policies, governance, and geographical regulations (different low regulation in different countries). A service provider gives users the access to resources using different authentication and authorisation roles and policies. They are enforced and monitored in equal interval. Such timing determinism may be exploited as a vulnerability to allow an attacker to use CPU time of the system in unauthorized way. The payment for CPU time will be at the expense of the end-user or system vendor. Putting unauthorized task between monitoring intervals will not be detected. Such vulnerabilities have been revealed during this timing-based manipulation [23]. There is a necessity of introducing non-deterministic methods during scheduling the tasks.

In this work, we define the problem of scheduling, monitoring, and reporting of independent tasks execution, where submitted tasks are processed in a batch mode using non-deterministic time intervals. Such process of managing tasks (scheduling, monitoring and reporting) is supported by a simple cloud multi-agent system.

In many existing security-driven scheduling models, service providers have full knowledge about the security demands from cloud users and the trust levels for virtual machines. Differently to the above approaches, in our model, we define dynamic “reputation” indexes for VMs and security requirements for the execution of tasks [10]. The security aspect of scheduling in our model can be interpreted as the system security concept from the provider’s perspective.

The paper is organised as follows. In section II, the model of Agent Supported Non-Deterministic Meta Scheduler is formally defined. The next two sections, III and IV, are dedicated to the implementation and evaluation of the proposed model respectively. The paper is summarized in Section V along with the directions of further research.

II. MODEL

We consider the IBS problem, where tasks are executed independently in batches. Additionally, we assume in our model that the times of the distribution of the workloads are stochastically estimated. Such scheduling problem instance can be defined by using $A|B|C$ notation specified in [8] and [12], where A defines the resource layer and architecture type, B defines the processing characteristics and the constraints, and C specifies the scheduling criteria. Formally, the considered scheduling problem can be defined as follows:

$$Rm|b, indep, stat, hier|(objectives) \quad (1)$$

where the individual components mean:

- Rm – tasks are sent into parallel resources of various computing capabilities;
- b - the task processing mode is batch mode;
- $indep$ - independency as the task interrelation;

- $stat$ - static mode, when the given number and characteristics of VMs remains the same during scheduling process;
- $hier$ - references that the scheduling objectives are optimised in hierarchical mode: a central meta-scheduler which interacts with local job schedulers in order to define the optimal schedules; and,
- $objectives$ - denotes the set of considered scheduling objective functions.

Conventional scheduling objectives are:

- The Makespan measures the time when the latest task is done:

$$C_{\max} = \min_{S \in Schedules} \left\{ \max_{j \in Tasks} C_j \right\}, \quad (2)$$

where C_j is the time when task j is finalized, $Tasks$ are all tasks submitted to system and $Schedules$ is the set of all possible schedules.

- The Flowtime measures minimum of the sum of times of all the tasks:

$$F = \min_{S \in Schedules} \left\{ \sum_{j \in Tasks} C_j \right\}, \quad (3)$$

A. Assumptions

Our scheduling model is based on a simple centralized multi-agent system with one master agent and several working agents (workers). The master agent manages the non-deterministic of tasks scheduling and execution. Then gathers and analyses the information about the objectives for the whole system [see (2)-(3)]. After that makes decisions about the scheduling process. Working agents look after the incorporated workload, and gather and analyse objectives for the nodes that they are responsible of. They also negotiate the parameters for the scheduling policy with the master agent.

The following conditions are necessary for definition and implementation of our scheduler:

- a fully distributed environment with M nodes composed of a VM;
- each task is performed on single VM;
- all tasks are delivered to the system in batches;
- a variety of computing capabilities, access modes, and response times for all participating nodes;
- the performance of any given task on a single VM is neither related nor affected by the performance of any other VM;
- the number of machines in the physical layer of the cloud is fixed and unchanged during the execution of generated schedules; and
- the scheduler sends the tasks from one batch after another.

Let the time interval $[0, T]$ be a time horizon of system activity. We assume that at time $t = 0$ the system receives a batch of tasks from the scheduler. Then the next one, until the

last batch. Last batch from comes before deadline time $t = T$. Let us assume that the scheduler is activated N times during time interval $[0, T], T > 0$, at randomly selected moments t_1, t_2, \dots, t_N - 'scheduler ticks'. At time $t = t_1$, the first batch of the tasks is sent and the scheduler waits until $t = t_2$. At $t = t_2$, some tasks will be completed and some will be in progress. There will be machines idle, busy with work to end the job, and some finishing their jobs. At $t = t_2$, the second batch of tasks is sent to the machines. Let us denote by t_i^{sched} the time that is necessary for the scheduler to perform all the activities to obtain the i -th schedule.

Let j from 1 to W (where W - number of tasks) be the task label. Task j may be characterised by workload parameter wl_j expressed in Millions of Instructions per second (MIPS). Then the vector $[wl_1, \dots, wl_W]$ is a workload vector for all tasks in the batch.

If we divide workload vector into K categories representing the K different levels task volume, we may build a frequency distribution, [13], that shows us a summarised grouping of tasks divided into mutually exclusive classes and the number of occurrences in a class. Firstly, we decide about the number of classes. The maximum number of classes may be determined by formula $\sqrt[3]{W}$. Secondly, we calculate the range of the data by finding minimum and maximum data value. The range will be used to determine the class width. Thirdly, we decide about width of the class denote by h and obtained by dividing the range by number of classes. The classes taken together have to cover the distance from the lowest value in the workload vector to the highest workload vector value. Finally, we find the frequencies in each class counting how many task belongs to each class. We formulate the frequency distribution by frequencies for all of the classes from the lowest values to the maximal values.

For example for $K = 3$, we may interpret categories as: low cost tasks, medium cost task, high cost task. The first (from three consecutive values of the frequency vector) coordinate denoted how many task is in the batch that require low cost to be performed.

The proposed model can be characterised as centralized adaptive IBS. Such schedulers can be represented by the vectors of machines or tasks labels where a direct representation of the schedule is used [14]. Let us denote by S the set of all permutations with repetition of the length W over the set of machine labels M . An element s from set S is termed as schedule and it is encoded by the vector $s = [i_1, \dots, i_W]$, where i_j from M denotes the number of the machine on which the task labelled by j is executed. The j from 1 to W are the task labels.

B. Non-deterministic ticks of scheduling/monitoring/reporting

Let us assume that the scheduling/monitoring/reporting is activated N times during time interval $[0, T], T > 0$, at random points t_1, t_2, \dots, t_N called ticks. Three different models of randomness were proposed to obtain t_i for $i = 1, 2, \dots, N$:

- discrete uniform distribution on interval $[0, T]$: Values are equally probable; every one of N values has equal probability $1/N$.

- Binary Blum-Blum-Shub sequence model: As shown in algorithm 1, the sequence of zeros and ones using Blum-Blum-Shub Pseudo-Random Bit Generator (PRBG) is calculated until the N -th ones appeared [18]. Then the interval is divided into K equal parts, where K - number of bits. The each subsequent value in the binary sequence is assigned to consecutive point. If the value zero was assigned to the particular tick - system remains turned off, if the value of one appeared scheduler/monitor/reporter is being turned on. The cryptographic security of the Blum-Blum-Shub PRBG is based on the quadratic residuosity problem [19].

Algorithm 1 Binary Blum-Blum-Shub sequence model

- 1: Generate p and q - two big Blum prime numbers.
 - 2: Calculate $n := pq$
 - 3: Choose the random seed $s \ni s \in (1, n - 1)$,
 - 4: Calculate $x_0 := s^2 \pmod n$
 - 5: **for** $i \in 1$ to n **do**
 - 6: $x_i := x_{i-1}^2 \pmod n$
 - 7: $z_i := x_i \pmod 2$
 - 8: **return** $z_1, z_2, z_3, \dots, z_n$
-

- Binary SHA-2 sequence model assumes the usage of cryptographic hash function Secure Hash Algorithm 2 in the SHA-512 version [2], with output size of 512 bits. That is the improvement of usage the older version of SHA-1 algorithm as the random number generator presented in [1]. This model assumes generation 512 values (0 or 1) at each single run and the scheme of turning on scheduler/monitor/reporter remains the same as in 3.

Both Pseudo-Random Bit Generators [18] are deterministic algorithms which generated binary sequence indistinguishable from a truly-random binary sequence. Therefore, it is impossible for the malicious individual to calculate when the next scheduling/monitoring/reporting will take place and the attack described in [23] is very probable to be revealed and blocked.

III. IMPLEMENTATION

The system is implemented using agents which are capable of autonomous actions in a given environment to meet design objectives. Specifically, the schedules are executed in non-deterministic intervals and monitored by dedicated agents. Moreover, the true response time of the nodes, and the time of processing the single task may vary non-deterministically due to the possible failures of the VM, being under denial-of-service attack, or problems with connection between machines.

In proposed model, the tasks are submitted to selected cloud task providers, from where they are sent to the pool of tasks. Next, the tasks are gathering to the batches. The master agent's inputs are both, batch of tasks and scheduling tick. The master agent is responsible for management of the scheduling process and decides when to get the batch and start scheduling. For the process of implementing of the schedule are responsible cloud task dispatchers, that submit the tasks to the VMs. The second amenability of master agent is the

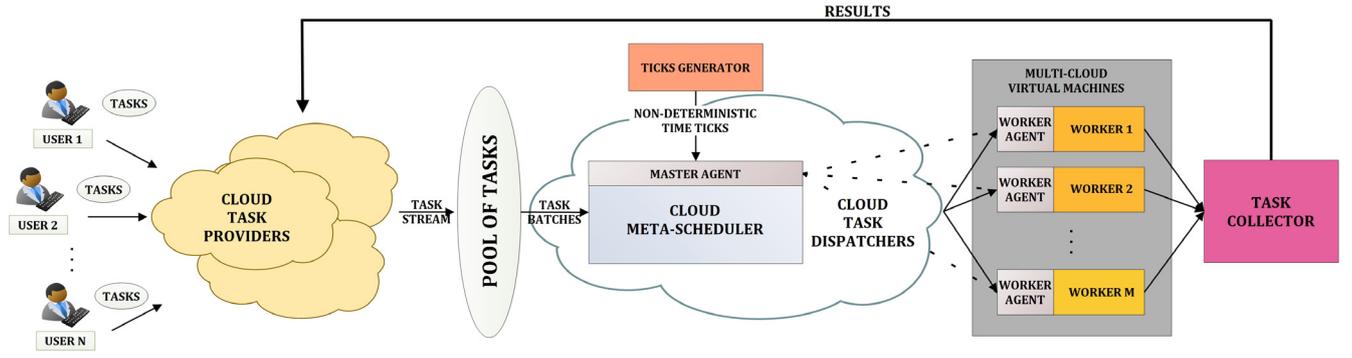


Fig. 1. Overall architecture of the system.

gathering the information from worker nodes agents, and making decisions respecting the their needs.

If the master agent is idle, it has to assume that the environment will not be at the same state, because the worker nodes are still in use. Thence, it has contact the worker agents to gather the information about the environment. This action is taken from time to time due to the very high cost of the continuous monitoring. The model has assumed that the activities of scheduling, monitoring, and reporting are being performed in the non-deterministic time intervals as the part of one connection.

Agents are assumed to have a set of possible actions, which transform the state of the environment. The architecture is shown in Figure 1.

Let $Ac = a_1, a_2, \dots, a_L$ be the (finite) set of L actions. The basic model of agents interacting with their environments assumes that the agents starts making decisions in some state of environment. The result of this decision that is the environment can respond with a number of possible states is $St = s_1, s_2, \dots$

This process has two main features. Firstly, the environment is history dependent. That means, the decisions made earlier by the agent influences the current state. Secondly, the environment is non-deterministic [21]. In this paper the we have proposed:

- a_i - action after setting non-deterministic i -th model, for $i = 1, 2, 3$ (see p. 6.),
- a_4 - action necessary to change between models 6.1-6.3,
- a_i - actions necessary to change parameters of models 6.1-6.3, for $i = 5, 6, 7, 8$,
- s_i - value of the chosen objective (2)-(3) for the whole system.

A utility function is defined as a numeric value, representing how 'good' the state is. The role of the agent is then to make such a decisions that maximize utility [21]. In this approach, utility is a function:

$$u : St \rightarrow R \quad (4)$$

which associates a real value with every environment state, and is defined as one of the possible objectives in equations (2)-(3).

The scheduling decisions influence the worker nodes. Additionally, the information from the worker nodes about the performance of the systems affect next Master Agents decision.

Each worker node in equipped with model of an agent responsible for monitoring the node and managing the tasks send to this node.

- a_i - action necessary to change between models 6.1-6.4, for $i = 1, 2, 3, 4$,
- a_i - actions necessary to change parameters of models 6.1-6.4, for $i = 5, 6, 7, 8$,
- a_i - actions after the k -th task arrived $i = 8 + k, \dots$,
- s_i - value of the chosen objective (2)-(3) for the particular node.

A FastFlow task farm has been selected as the testing environment. The task-farm pattern allows for the replication of a stateless function W (Worker), where each replica receives stream inputs from a dispatcher E (Emitter) and sends outputs to the collector C . The master and worker agents have been deployed as separate services in the form of private functions for workers and FastFlow Emitters.

FastFlow is a structured parallel programming environment written in C++ on top of the POSIX thread library. It provides programmers with predefined and customisable stream parallel patterns—including task-farm and pipeline—and has been successfully employed in clouds [6].

The implementations of Blum-Blum-Shub Pseudo-random number generator and SHA-2 based Pseudo-random number generator have been programmed via the OpenSSL C++ library [20]. The pseudo-random number generation uniform distributions employs the Random Number Engines and Random Number Distributions for C++. These algorithms are the basis for the ticks generator.

IV. EVALUATION

The empirical evaluation is based on the matrix multiplication problem and the size of the task governed by the magnitude of the matrix. Five types of tasks were considered. The ratio of the tasks size in term of time consumption was: 1, 2, 3, 8, 10, where the size 1 is equal to single node working for $T/12$ [min] that is 5 minutes. 120, 180 or 240 tasks are grouped into a single batch. Each batch is assumed to be the

same, but the single tasks are scheduled using round-robin according to the scheduler policy.

Five types of workers are introduced. For the sake of clearness workers was arranged as follows:

- Very high speed worker A, using 100% of the computational time available in favour of the system;
- High speed worker B, using 90% of the computational time available in favour of the system;
- Moderate speed worker C using 75%;
- Slow speed worker D using 60%;
- Very slow worker E using 50%, according the introduced referential time slice.

This simulation reflects the different worker physical location and the variety of time necessary to connect with the worked, sending the data to be preprocessed it and gathering the results of calculation. Measures of performance of the system selected were: makespan, see (2) and flowtime, see (3).

Three types of load have been considered:

- Low: workers remains idle for some amount of given time;
- Moderate: workers are idle for very small amount of the total time;
- High: workers do not get idle at all.

For comparative purposes, a referencing model has also been introduced. In this model, the time interval $[0, T]$ was divided into p equal parts, where p was assumed to be equal to 12. The batch of tasks is distributed 12 times by the round robin scheduler. It has been presumed during models 6.1 – 6.3 testing that the same number of scheduler switching on (and the number of the system checks and reports) introduced is 12.

Consequently, we have three scenarios: 120, 180 and 240 tasks distributed during time interval $[0, T]$. The time load of each batch of task was (according to a very high speed worker computational power, not considering lateness due to communication and other reasons): 5 tasks that lasts $(1/5) * (T/12)$, 5 tasks that lasts $(2/5) * (T/12)$, 5 tasks that lasts $(3/5) * (T/12)$, 5 tasks that lasts $(1+3/5) * (T/12)$, 5 tasks that lasts $2 * (T/12)$. Each batch has the same size and the workload is constant in order to check the influence of the models 6.1-6.3 into the objectives (2)-(3). For the clarity of the presentation T is assumed as 3600 seconds. Models 6.1-6.3 resulted in different time moments in with the tasks were send to the workers.

TABLE I. THE FREQUENCY OF WORKERS A, B, C, D, E USAGE DURING PERFORMING 12 BATCHES X 20 TASKS=240 TASKS

Model	A	B	C	D	E
Ref.	65	53	45	42	35
Uniform	55	56	48	41	40
BBS	63	51	51	38	37
SHA-2	61	51	50	39	39

The utilization of the workers by the models is shown in the Table I. The Uniform model has used the fastest workers

significantly more seldom, as they were used by the Referencing Model, BBS, and SHA-2 models. The best resources usage seem to take place in the Referencing Model but BBS and SHA-2 models performed similarly and not worse.

TABLE II. MAKESPAN EQ (2) VALUE [SEC] FOR DIFFERENT NUMBER OF TASKS

Model	120	180	240
Ref.	14185 (3,9h)	14523 (4,3h)	17161 (4,7h)
Uniform	10522 (2,9h)	13260 (3,6h)	17341 (4,8h)
BBS	12921 (3,5h)	13959 (3,8h)	18650(5,1h)
SHA-2	13124 (3,6h)	14018 (3,8h)	17610 (4,8h)

Makespan eq(2) value has been examined for different system loads as shown in Table II. The big potential for upgrading the system performance without intervention inside the scheduler is stated. Changing the time of scheduler usage lead to the reduction of the Makespan of the given tasks for all non-deterministic models for low to moderate system load. When the system was very heavy loaded the models seem very close as far as the Makespan of the given tasks.

TABLE III. MEAN MAKESPAN AND STANDARD DEVIATION FOR 12 BATCHES FOR VARIOUS NUMBERS OF TASKS IN THE BATCH

nr of tasks in batch	mean	st. dev.
Ref. 10	2188	353
Ref. 15	2265	367
Ref. 20	2411	214
Uniform 10	1796	215
Uniform 15	2120	256
Uniform 20	2454	284
SHA-2 10	1705	193
SHA-2 15	2115	334
SHA-2 20	2383	266
BBS 10	1767	185
BBS 15	2100	314
BBS 20	2484	526

Table III presents the mean Makespan of 12 batches, where all batches are equal in terms of workload. The differences in Makespans comes from the necessity of waiting in the queue for the worker to be available. The test shows that for all models, the mean Makespan is kept at constant level with a non-decreasing standard deviation.

The higher value of standard deviation of Makespan value means that some bottlenecks occurred during the calculation process. The raising standard deviation may be the the signal for the worker agents to start the negotiation with the master agent to switch from one model to another. Alternating the models may help the master agent to choose the most effective one.

For BBS and SHA-2 models it is easy to calculate the schedules more often without the need to run the model again, and the next part of the 0-1 sequence may be used with different frequency. The assumed frequency might be changed from one minute to 0.5 min or any given number of seconds. For the discrete uniform model 6.1 is necessary to prepare new set of random numbers.

The sample distribution of time ticks 'worst case scenarios' (regardless of non-deterministic values, first scheduling is done at the beginning of the process for the system not to remain idle) may be:

- 1 31 40 48 50 25 37 14 24 38 35 51 from the discrete uniform model 6.1, that gives the time of scheduling during time interval length $T=60\text{min}$: 1min, 14min, 24min, 25min, 31min, 37min, 38min, 36min, 40min, 48min, 50min, 51min. It results 4 times of scheduling/monitoring in the first half if the time interval, and 7 times of scheduling/monitoring in the second half if the time interval. This kind of scheduling adversely affects the time of calculation if the batches are so small that workers become idle waiting for the next batch, after finishing all of the tasks from the previous batch. When workers are very loaded, this way of scheduling does not affect the performance of the system much, because during the next scheduling tasks are queued in workers that are still in use.
- 010111111001001101: 12-ths 'one' as 19-ths number inside the sequence of BBS 6.2 model, that gives the time of scheduling during time interval length $T=60\text{min}$: 1min, 3min, 9min, 12min, 15min, 18min, 21min, 24min, 27min, 36min, 45min, 48min, 54min. The assumed frequency in minutes. It results 9 times of scheduling/monitoring in the first half if the time interval, and 4 times of scheduling/monitoring in the second half if the time interval.
- 111110011011001100000000001: 12-ths 'one' as 28-ths number inside the sequence of SHA-2 6.3 model, that gives the time of scheduling during time interval length $T=60\text{min}$: 1min, 4min, 6min, 8min, 10min, 16min, 18min, 22min, 24min, 30min, 32min, 58min. It results 9 times of scheduling/monitoring in the first half if the time interval, and 3 scheduling times in the second half if the time interval. The phenomenon mentioned above is taking place in the second half of the $[0,T]$ interval.
- the time of scheduling during time interval length $T=60\text{ min}$ in the reference deterministic model: 1min, 5min, 10min, 15min, 20min, 25min, 30min, 35min, 40min, 45min, 50min, 55min. It results in 6 times of scheduling/monitoring in the first half if the time interval, and 6 scheduling times in the second half if the time interval.

The sets of numbers presented above are the single execution of the proposed models. For the process of scheduling that lasts much longer, e.g. one month, we will obtain different sets with distinct possible 'gaps' between scheduling/monitoring/reporting moments. The average behaviour of the model after many executions will not depend much on the single realization's imperfection.

V. CONCLUSIONS AND FUTURE WORK

The proposed models for modifying the existing schedules policies have proven to be effective. They enable fluent batch calculating without any side effects. The non-deterministic intervals of system monitoring and reporting prevents from timing attacks on the resources. Different moments of sending batches to Virtual Machines enable to improve system speed as far as the Makespan is concerned.

Our case study with corresponding experimental results has demonstrated the effectiveness of the proposed solution and

the potential to increasing the system effectiveness without changing the scheduler itself.

The proposed models have checked resources consumption smoothness. Makespan of the pool of task was examined, the average time of single batch processing and the bottlenecks occurrence. There are differences accordingly to the system load. The preliminary strategies of master and worker agents have been developed. The proposed multi-agent model enables to negotiate the parameters of the models according to the system state.

The proposed solution is very elastic and may be used for different scheduler types. This approach did ensure proper security of the system. Monitoring and reporting are not possible to predict, because of the fact that they are based on 'safe' random number generator (BBS) or generator working as the random oracle (SHA).

Achievement of the aforementioned objectives open several possibilities for further research in related areas. Thus, the number of additional tests and extensions are planned in the nearest future. One of them is to consider the non-heterogeneous batches of tasks. Based of the knowledge of the length of the next gap between scheduler usage, it is planned to incorporate the master agent changing decision about the enlargement of the batch size. Worker agents are planned to negotiate the smallest gaps when they are recording that idle state occurred.

The Artificial Neural Network (ANN) is being tested as the support for the master agent decisions. ANN is simulating the cloud computing environment without running the working nodes. It calculates the estimate results of the different future decision of the master agent, helping it to proceed towards better option.

The project involves the implementation of new non-deterministic models for ticks of scheduling, monitoring and reporting. One of these models can be discrete Poisson distribution. As well as plans to implement new schedulers - especially those based on genetic algorithms. More advanced schedulers will be tested in OpenStack and Amazon AWS environments.

ACKNOWLEDGEMENT

The inspiration for the presented research was the work of Agnieszka Jakóbk and Daniel Grzonka under supervision of Joanna Kołodziej and Horacio Gonzalez-Velez at National College of Ireland (NCI) during the short scientific visit in Dublin (Ireland) according to the STSM programme. The STSM programme is supported by the ICT COST Action IC1406 (cHiPSet) "High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)".

REFERENCES

- [1] Digital signature standard (DDS). Technical report, 2013.
- [2] Secure hash standard (SHS). Technical report, 2015.
- [3] H. Arabnejad and J. G. Barbosa. List scheduling algorithm for heterogeneous systems by an optimistic cost table. *Parallel and Distributed Systems, IEEE Transactions on*, 25(3):682–694, 2014.
- [4] M. J. Atallah and M. Blanton. *Algorithms and Theory of Computation Handbook, Volume 2: Special Topics and Techniques*. CRC press, 2009.

- [5] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic. Cloud computing and emerging {IT} platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6):599–616, 2009.
- [6] S. Campa, M. Danelutto, M. Goli, H. González-Vélez, A. M. Popescu, and M. Torquati. Parallel patterns for heterogeneous CPU/GPU architectures: Structured parallelism from cluster to cloud. *Future Generation Computer Systems*, 37:354–366, 2014.
- [7] H. M. Chaskar and U. Madhow. Fair scheduling with tunable latency: a round-robin approach. *IEEE/ACM Transactions on Networking (TON)*, 11(4):592–601, 2003.
- [8] P. Fibich, L. Matyska, H. Rudová, et al. Model of grid scheduling problem. *Exploring Planning and Scheduling for Web Services, Grid and Autonomic Computing*, pages 17–24, 2005.
- [9] D. Ge, Z. Ding, and H. Ji. A task scheduling strategy based on weighted round robin for distributed crawler. *Concurrency and Computation: Practice and Experience*, 2015. In press. DOI: 10.1002/cpe.3701.
- [10] D. Grzonka, J. Kołodziej, J. Tao, and S. U. Khan. Artificial neural network support to monitoring of the evolutionary driven security aware scheduling in computational distributed environments. *Future Generation Computer Systems*, 51:72–86, 2015.
- [11] D. Karger, C. Stein, and J. Wein. Scheduling algorithms. In *Algorithms and theory of computation handbook*, pages 20–20. Chapman & Hall/CRC, 2010.
- [12] D. Klusáček and H. Rudová. Efficient grid scheduling through the incremental schedule-based approach. *Computational Intelligence*, 27(1):4–22, 2011.
- [13] O. Knill. *Probability and stochastic processes with applications*. Overseas Press, 1994.
- [14] J. Kołodziej. *Evolutionary Hierarchical Multi-Criteria Metaheuristics for Scheduling in Large-Scale Grid Systems*, volume 419. Springer, 2012.
- [15] J. Kołodziej and S. U. Khan. Multi-level hierarchic genetic-based scheduling of independent jobs in dynamic heterogeneous grid environment. *Information Sciences*, 214:1–19, 2012.
- [16] P. M. Mell and T. Grance. The nist definition of cloud computing. sp 800-145. Technical report, 2011.
- [17] N. K. Rajput and A. Kumar. A task set based adaptive round robin (tarr) scheduling algorithm for improving performance. In *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on*, pages 347–352. IEEE, 2015.
- [18] A. Sidorenko and B. Schoenmakers. Concrete security of the blum-blum-shub pseudorandom generator. In *Cryptography and Coding*, pages 355–375. Springer, 2005.
- [19] D. R. Stinson. *Cryptography: theory and practice*. CRC press, 2005.
- [20] J. Viega, M. Messier, and P. Chandra. *Network Security with OpenSSL: Cryptography for Secure Communications*. O'Reilly Media, Inc., Sebastopol, 2002.
- [21] M. Wooldridge, N. R. Jennings, et al. Intelligent agents: Theory and practice. *Knowledge engineering review*, 10(2):115–152, 1995.
- [22] X. Yuan and Z. Duan. Fair round-robin: A low complexity packet scheduler with proportional and worst-case fairness. *Computers, IEEE Transactions on*, 58(3):365–379, 2009.
- [23] F. Zhou, M. Goel, P. Desnoyers, and R. Sundaram. Scheduler vulnerabilities and coordinated attacks in cloud computing. *Journal of Computer Security*, 21(4):533–559, 2013.

AUTHOR BIOGRAPHIES



AGNIESZKA JAKÓBIK (KROK) received her M.Sc. in the field of stochastic processes at the Jagiellonian University, Cracow, Poland and Ph.D. degree in the field of neural networks at Tadeusz Kosciuszko Cracow University of Technology, Poland, in 2003 and 2007, respectively. From 2009 she is an Assistant Professor at Faculty of Physics, Mathematics and Computer Science, Tadeusz Kosciuszko Cracow University

of Technology. Her main scientific and didactic interests are focused mainly on Artificial Intelligence: Artificial Neural Networks, Genetic Algorithms, and additionally on Parallel Processing and Cryptography. Her e-mail address is: agneskrok@gmail.com



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. Currently, he is Research and Teaching Assistant at Cracow University of Technology and Ph.D. student at Jagiellonian University in cooperation with Polish Academy of Sciences. He is also a member of Polish Information Processing Society and IPC member of several international conferences. The main topics of his research are grid and cloud computing, multi-agent systems and high-performance computing. For more information, please visit: www.grzonka.eu



HORACIO GONZÁLEZ-VÉLEZ is currently an Associate Professor and Head of the Cloud Competency Centre at the National College of Ireland. He spent over a decade working in systems engineering and product marketing for innovation-driven companies such as Silicon Graphics and Sun Microsystems, he earned a PhD in Informatics from the University of Edinburgh. E-mail: horacio@ncirl.ie



JOANNA KOŁODZIEJ is an associate professor in Department of Computer Science of Cracow University of Technology. She is a vice Head of the Department for Sciences and Development. She serves also as the President of the Polish Chapter of IEEE Computational Intelligence Society. She published over 150 papers in the international journals and conference proceedings. She is also a Honorary Chair of the HIPMOS track of ECMS. The main topics of here research is artificial Intelligence, grid and cloud computing, multiagent systems. The detailed information is available at www.joannakołodziej.org

THE MODEL OF DATA DELIVERY FROM THE WIRELESS BODY AREA NETWORK TO THE CLOUD SERVER WITH THE USE OF UNMANNED AERIAL VEHICLES

Ruslan Kirichek

Department of Telecommunication Networks and Data Transmission
State University of Telecommunication,
193232, St.Petersburg, Russia
E-mail: kirichek@sut.ru

KEYWORDS

Model, WBAN, UAV, Data, Sensor Node, Channel.

ABSTRACT

Currently, the wireless body area network for monitoring the human body and preventing the critical situations of life are widely spread. Data from the sensors are collected in the memory of the microcomputer, and should be delivered in a cloud service for the further analysis. The classical methods for delivery via wireless interfaces require large power consumption (Wifi, LTE, 3G technology), whereas the protocols with low power consumption suggest the presence of additional gateway with the Internet, which is not always possible to implement. The paper proposes an original approach of using UAVs to deliver data from the wireless body area network to a remote cloud server. The proposed model takes into account the bandwidth, the ability of generating traffic parameters, the number of sensor nodes and the characteristics of the UAV.

INTRODUCTION

The Flying Ubiquitous Sensor Network (FUSN) is a new application of the Internet of Things (Koucheryavy et al. 2015; Kirichek and Koucheryavy 2016). The term is rather new, but it has managed to win the attention of the scientific community. The main objective of the FUSN is a collection of data from the remote sensor nodes by using public unmanned aerial vehicles (UAV-P). This effect has become possible and necessary due to the emergence of a large number of objects, equipped with sensors (temperature sensors, humidity, light, etc.) and radiomodules based on protocols: ZigBee, 6LoWPAN, BLE, RPL, AODV and so on (Koucheryavy et al. 2014; Vasiliev et al. 2014; Rosario et al. 2014), as well as a reduction of the cost and promotion the UAV-P that can perform the autonomous flights on a given route (Phuong et al. 2014; Kirichek et al. 2015).

There are two segments of the flying ubiquitous sensor networks: terrestrial and flying. The terrestrial segment is represented as a classical ubiquitous sensor network (Abakumov and Koucheryavy 2014; Vybornova and Koucheryavy 2014; Futahi et al. 2015), and the flying segment as one or several of UAV-P (Sahingoz 2014; De Freitas et al. 2010; Orfanus et al. 2014). Actively

developing medical wireless body area networks (WBAN) are (Kirichek et al. 2016):

- Wearable smart fitness electronics, which acts as a mentor in the observance of a healthy way of life (sleep, physical activity, etc.). The data is collected from all kinds of bracelets, smart watches, pedometers, neurointerface and then is delivered to the cloud service through the Internet, while the smartphone is used as a gateway. As a rule, these products do not require strict medical certification.

- Specialized clothing and high-precision sensors for collecting the data on the state of human health data according to the key performance indicators (heart rate, blood sugar, blood pressure and others.). Sensors are installed in the suits of sportsmen and people of extreme professions. These products are subject to mandatory medical certification.

- Medical nanonetworks are implanted directly into the body of the human and communicate through a dedicated gateway. This area is at the stage of research, but the results are predicting the widespread of this type of network.

In turn, the collected data should be delivered to a single point for storage and subsequent processing. For these purposes, it is advisable to use UAV-P that can quickly fly around the area of people dislocation and transmit data to the cloud server, using the telemetry channel.

An example of such an interaction can be a group of climbers who came under an avalanche of snow. UAV-P is capable of arriving at the accident site in a few minutes, flying around the area, fixing the place where there are people and transmitting the data from the WBAN about their health in the cloud server.

OVERVIEW OF THE UAV CHANNELS OF INTERACTION WITH THE PUBLIC COMMUNICATIONS NETWORK

One or more UAV can be used for the tasks of data collection and transmission (figure 1). Autonomous UAV flight is performed on the basis of the flight task, which is pre-loaded into the controller (Kirichek et al. 2015).

Although the full autonomy of the flight, the telemetry channel is typically used for monitoring parameters of UAV and for supplying the emergency landing command. Given the fact that the maintenance traffic is

transmitted in the channel – the channel utilization is about 60%, the channel resource remains small for efficient data transmission collected from WBAN.

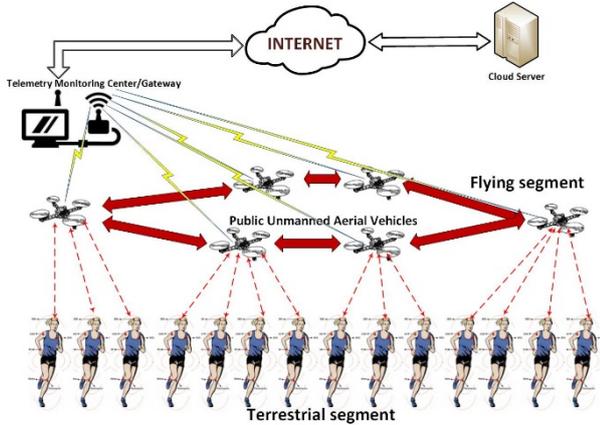


Figure 1: Data Collection from WBAN Using One or Multiple UAVs

Thus, it is necessary to solve the problem of guaranteed data delivery from the sensor nodes (a man with WBAN) on the cloud server via the Internet.

The following parameters can be varied:

- The number of people;
- The number of UAVs;
- The radius of radio coverage of the UAV;
- The size of the territory with people;
- Frequency (various channels of interaction).

THE MODEL OF DATA DELIVERY FROM THE WBAN TO THE PUBLIC COMMUNICATION NETWORK WITH THE USE OF UAV

The purpose of the UAV network simulation is to choose its parameters to provide the desired quality of service traffic (Kirichek et al. 2015). The quality of service traffic in the context of this task is determined by the delivery of the data from the sensor network (WBAN) in the cloud server. The network parameters are the bandwidth data transmission channel and the number of UAVs required for the task.

In order to build a model, we will make the following assumptions. Let us assume that it is required to transfer the data from a group of n sensor nodes. The sensor nodes are arranged randomly and form a Poisson field (Ventcel' 1959). Each of the sensor nodes produces a fixed amount of data v (bytes), which is a set of values obtained from the various sensors. The data from each of the nodes should be delivered to the collection cloud server at a time not exceeding T_0 . The network structure includes a segment of the sensor nodes, the access network segment, implemented with the help of UAVs, gateway and data network segment between the gateway and the cloud services server. We assume that the bottleneck in this structure is a segment of the access network. Further we will consider only the settings for the access network segment, believing that the network settings of other elements have a substantial margin of bandwidth.

The access network is constructed on the basis of the communication nodes located on UAVs. In general, k UAVs can be located in a service area, where the sensor nodes coverage areas may overlap, as shown in figure 2, 3.

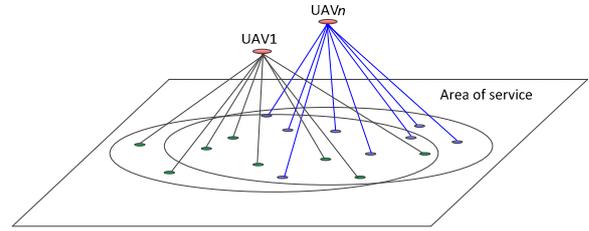


Figure 2: The Sensor Network Areas Model Served by UAVs

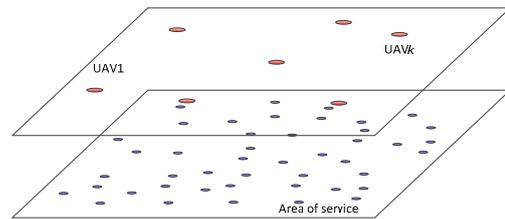


Figure 3: Overlap Sensor Network Areas and UAVs Areas

Area can be served by multiple UAVs. UAVs have a random location at some time, and can also be viewed as a Poisson field.

We believe that the communication nodes which are located on UAVs, have a choice of free frequency channel to communicate with the sensor nodes. Frequency Division Multiplexing makes it possible to serve the nodes of sensor network with multiple UAVs crossing their service areas. We assume that the sensor node periodically performs the data sending operation. If at the time of this operation there is a free channel resource, the data is transmitted at a rate determined by the free channel bandwidth. If there is no available resource, the data waits to be sent. Time data delivery through the access network will be determined by the transmission rate in the channel and the standby transmission start time

$$\bar{T} = \bar{W} + \bar{t} \quad (1)$$

where \bar{W} - waiting time of the beginning of transfer; \bar{t} - transfer time through a channel.

Considering this system as a queuing system, it is possible to estimate the delay associated with waiting for the start of service.

The flow of requests for the system input is determined by the activity and the number of the sensor nodes. We proceed from the fact that the traffic characteristics, which are produced by various different nodes, differ (different frequency of sending data and different amount of data). Then the resulting flow of quite a large number of nodes can be described by a simple flow model, the intensity of which is equal to

$$\lambda = \sum_{i=1}^n \lambda_i \quad (\text{requests/sec}) \quad (2)$$

where λ_i – the flow rate produced by the i -th node. The average volume of transmitted data is:

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i \quad (\text{bit}) \quad (3)$$

where v_i is the volume of data transmitted by the i -th node.

At a constant speed of data transmission in the channel b (bit/s) the average service time of requests is equal to

$$\bar{t} = t_c + \frac{\bar{v}}{b} \quad (\text{s}) \quad (4)$$

where t_c - the time required for channel selection and communication (sec).

v - the amount of sensor data (bit);

b - the data transmission rate in the channel (bits/sec).

If t_c is significantly less time required for transmission, it can be neglected.

If at the moment of awakening the node selects a free, at the moment, frequency channel, ie, channel which is not currently transmitted to a traffic, then the model can be described with a queuing system form M/G/k in service procedure with expectation. Assuming that the communication channels with the UAV $1 \dots k$, are identical, it is possible to present that their total capacity is equal to

$$\mu_k = \frac{k}{\bar{t}} \quad (\text{requests/sec}) \quad (5)$$

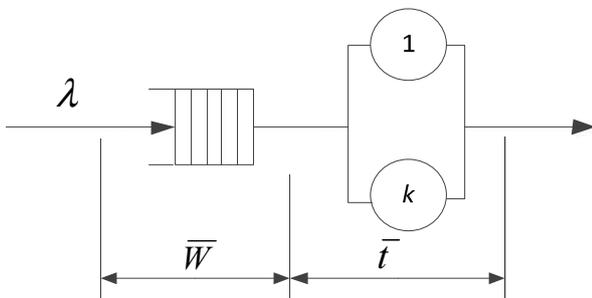


Figure 4: Model of Service Sensor Network Nodes

If we make the assumption that the time of service requests (the time of its transmission over the channel) \bar{t} is accidental and has an exponential distribution, for the system M/M/k we can estimate the waiting time at queue as the

$$\bar{W} = \frac{E_C(k, \lambda, \mu)}{k\mu - \lambda} \quad (6)$$

where $E_C(k, \lambda, \mu) = \frac{a^k \cdot k}{k! \cdot (k-a)} \cdot \frac{1}{\sum_{i=0}^{k-1} \frac{a^i}{i!} + \frac{a^k}{k!} \cdot \frac{k}{k-a}}$ is the

second formula Erlang (Erlang C-formula) (Iversen 2001);

$a = \lambda \cdot \bar{t}$ - the intensity of the load (Earl);

k - the number of available channels (UAV);

$\mu = \frac{1}{\bar{t}}$ - the average number of requests served (sent) by one channel per unit of time.

If the distribution of the service time is different from the exponential, it is possible to resort to an approximate evaluation of its upper limit, for example (Kleinrock 1975).

Assuming an exponential distribution service time, the necessary number of available channels (UAV) can be estimated as the solution of the inequality $\bar{T} \leq T_0$ that can be obtained by minimizing the expression

$$k = \arg \min_k \left| \frac{E_C(k, \lambda, \mu)}{k\mu - \lambda} + \bar{t} - T_0 \right| \quad (7)$$

The conditions of the model k - should be viewed as the average number of UAVs, which are simultaneously in communication area where the density can be obtained, i.e. the average number of UAV unit area service area

$$\rho_{UAV} = \frac{k}{\pi R^2} \quad (8)$$

where R – the radius of the node connection.

Thus, using the expressions (7) and (8) based on the intensity of traffic data which is produced by the sensor nodes, and their amount of traffic intensity, the amount of transmitted data and the radius of connection, one can estimate the required number (density) of the UAV which is necessary for data delivery over the time, which does not exceed a predetermined value.

CONCLUSION

The paper suggests the new ways of using UAVs to deliver the real-time data on a remote cloud server. The proposed model will enable to calculate the number of UAVs required for the collection and delivery of data, taking into account the intensity and volume of traffic to the network underwear, the number of units and the radius of their connection.

These results will be used to conduct a series of experiments on the flying ubiquitous sensor network.

ACKNOWLEDGMENT

The reported study was supported by RFBR, research project No.15 07-09431a “Development of the principles of construction and methods of self-organization for Flying Ubiquitous Sensor Networks”.

REFERENCES

- Abakumov, P.; and A. Koucheryavy. 2014. "The Cluster Head Selection Algorithm in the 3D USN". *Proceedings of the 16th International Conference on Advanced Communication Technology, ICACT 2014* (Phoenix Park, Korea, Feb. 16-19), IEEE, 462-466.
- De Freitas, E. P.; T. Heimfarth; I. F. Netto; C. E. Lino; C. E. Pereira; A. M. Ferreira; F. R. Wagner; and T. Larsson. 2010. "UAV relay network to support WSN connectivity". *Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems 2010* (Moscow, Russia, October 18-20), IEEE, 309-314.
- Futahi, A.; A. Koucheryavy; A. Paramonov; and A. Prokopiev. 2015. "Ubiquitous Sensor Networks in the Heterogeneous LTE Network". *Proceedings of the 17th International Conference on Advanced Communication Technology, ICACT 2015* (Phoenix Park, Korea, July. 1-3), IEEE, 28-32.
- Iversen, Villy B. 2001. *Handbook "Teletraffic engineering"*. ITU-D Study Group 2, Question 16/2, June 20, 2001
- Kirichek, R.; A. Paramonov; and K. Varedzhyan. 2015. "Optimization of the UAV-P's Motion Trajectory in Public Flying Ubiquitous Sensor Networks (FUSN-P)". *Proceedings of the 15th International Conference, NEW2AN 2015, and 8th Conference, ruSMART 2015* (St.Petersburg, Russia, August 26-28), LNCS 9247, Springer, 352-366.
- Kirichek, R.; A. Paramonov; and A. Koucheryavy. 2015. "Flying Ubiquitous Sensor Networks as a Quening System". *Proceedings of the 17th International Conference on Advanced Communication Technology, ICACT 2015* (Phoenix Park, Korea, July. 1-3), IEEE, 127-132.
- Kirichek, R.; and A. Koucheryavy. 2016. "Internet of Things Laboratory Test Bed". *Proceedings of the Wireless Communications, Networking and Applications 2014, LNEE*, Springer, T.348, 485-494.
- Kirichek, R.; R. Pirmagomedov; R. Glushakov; and A. Koucheryavy. 2016. "Live Substance in Cyberspace - Biodriver System". *Proceedings of the 18th International Conference on Advanced Communication Technology, ICACT 2016* (Phoenix Park, Korea, Jan.31-Feb.3), IEEE, 274-278.
- Kleinrock, L. 1975. "Queueing Systems". John Wiley & Sons.
- Koucheryavy, A.; A. Muthanna; A. Prokopiev; and A. Paramonov. 2014. "Comparison of protocols for Ubiquitous wireless sensor network". *Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops 2014* (St. Petersburg, Russia, October 6-8), IEEE, 434-437.
- Koucheryavy, A.; A. Vladyko; and R. Kirichek. 2015. "State of the Art and Research Challenges for Public Flying Ubiquitous Sensor Networks". *Proceedings of the 15th International Conference, NEW2AN 2015, and 8th Conference, ruSMART 2015* (St.Petersburg, Russia, August 26-28), LNCS 9247, Springer, 299-308.
- Orfanus, D.; F. Eliassen; and E.P. de Freitas. 2014. "Self-Organizing Relay Network Supporting Remotely Deployed Sensor Nodes in Military Operations". *Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops 2014* (St. Petersburg, Russia, October 6-8), IEEE, 326-333.
- Phuong, T. H.; H. Yamamoto; and K. Yamazaki. 2014. "Data Synchronization Method in DTN Sensor Network using Autonomous Air Vehicle". *Proceedings of the 16th International Conference on Advanced Communication Technology, ICACT 2014* (Phoenix Park, Korea, Feb. 16-19), IEEE, 382 - 387.
- Rosario, D.; Z. Zhao; T. Braun; E. Cerqueira; and A. Santos. 2014. "A Comparative Analysis of Beaconless Opportunistic Routing Protocols for Video Dissemination over Flying Ad-hoc Networks". *Proceedings of the 14th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems, NEW2AN 2014* (St.Petersburg, Russia, August 27-29), LNCS 8638, Springer, 253-265.
- Sahingoz, O. K. 2014. "Networking Model in Flying Ad Hoc Networks (FANETs): Concepts and Challenges". *Journal of Intelligent & Robotics Systems. V.74, issue 1-2*, Springer, 513-527.
- Vasiliev, D. S.; D. S. Meitis; and A. Abilov. 2014. "Simulation-Based Comparison of AODV, OLSR and HWMP Protocols for Flying Ad Hoc Networks". *Proceedings of the 14th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems, NEW2AN 2014* (St.Petersburg, Russia, August 27-29), LNCS 8638, Springer, 245-252.
- Ventcel', E.S. 1959. "Probability theory". M. "Nauka" (in Russian).
- Vybornova, A.; and A. Koucheryavy. 2014. "Traffic Analysis in Target Tracking Ubiquitous Sensor Networks". *Proceedings of the 14th International Conference on Internet of Things, Smart Spaces, and Next Generation Networks and Systems, NEW2AN 2014* (St.Petersburg, Russia, August 27-29), LNCS 8638, Springer, 389-398.

AUTHOR BIOGRAPHIE



RUSLAN KIRICHEK was born in 1982 in Tartu (Estonia). He graduated Military-Space Academy A.F. Mozhaiskogo and St.Petersburg University of Telecommunication in 2004 and 2007 respectively. R.Kirichek received Ph.D from St.Petersburg University of Telecommunication in 2012. Since 2004 he worked at IT-department of the Air Force as a senior engineer. Since 2008 worked as a senior researcher at the FSUE "Research and Development Center "Atlas". Supervised research testing communication networks in terms of destructive influences. Since 2012 worked as the Head of the Internet of Things Laboratory at St.Petersburg University of Telecommunication. His e-mail address is: kirichek@sut.ru and his Web-page can be found at <http://www.iotlab.ru>

SIMULATION OF ROBOT-ASSISTED WSN LOCALIZATION USING REAL-LIFE DATA

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

WSN; Wireless Sensor Networks; Localization; Robot movement; Real-life data

ABSTRACT

The paper concerns the problem of robot-assistance in localization of stationary nodes in Wireless Sensor Networks (WSN). In our work, we present a simulation of localization process based on real-life data obtained during experiments. The paper describes classical localization algorithm – multilateration, however this algorithm is run not only on raw distance estimates but also on distance estimates obtained after application of specialized filtering procedures. A provided case study demonstrates the localization accuracy obtained for a robot localizing three stationary nodes during its movement along example path.

INTRODUCTION

Localization of moving objects is one of the fastest developing topics in the field of wireless communication. The domain which few years ago was a point of interest mostly researchers working in robotics, now becomes more and more popular in other fields. The idea of providing information about mobile user position especially in indoor environment attracts attention not only researchers, but first of all a huge number of companies. They hope this technology will open for them a new advertisement market. As a confirmation one can look at the well known localization contest organized in recent years by Microsoft co-located with IPSN conference (International Conference on Information Processing in Sensor Networks). The best solutions published last year provided means and algorithms allowing to locate user with localization error lower than half of meter (see Dewberry and Petroff (2015); Sánchez et al. (2015); Lazik et al. (2015)).

The obtained accuracy is impressing. However there is one disadvantage – each of this solutions require highly specialized devices and techniques to achieve such accuracy. For example Sanchez et al. propose to build a map of the environment.

Their system employs a high definition point cloud generated by pair wise registration of laser scanner acquisitions, yielding accurate maps consisting of 3D points, normals and (optionally) colors. As a consequence this method can be applied only in environment where specialized mapping procedure was done earlier. Overcoming this drawback is difficult and usually influence worse accuracy, however the survey paper published by Liu et al. (2007) outlines at least few methods applying Received Signal Strength (RSS) measurements for localization of moving objects.

High localization quality obtained by our High Performance Localization System (HPLS) in case of static networks (see Marks et al. (2014)) encouraged us to validate by simulation the possibility of localizing moving object in WSN environments. The simulation is done numerically but is based on real-life data registered in existing network composed by set of Crossbow MicaZ nodes.

This paper is organized as follows. Section II explains all aspects of collecting and processing RSSI data from testbed networks. Section III introduces the localization task formulation. Next this task is solved using localization scheme described in Section IV. Section V presents some numerical results obtained in our test network. Finally, Section VI concludes the paper and gives possible future directions for research on robot-assisted wireless positioning.

REAL-LIFE RSSI DATA

Received Signal Strength Indicator (RSSI) is considered to be a simplest and cheapest method amongst the wireless distance estimation techniques, since it does not require additional hardware for distance measurements and is unlikely to significantly impact local power consumption, sensor size and thus cost. Main problem in application RSSI is low accuracy. According to well-known wireless channel models (described in next section) received power should be a function of distance. However, the RSSI values have a high variability and it's difficult to use them as a distance estimator (see Benkic et al. (2008); Ramadurai and Sichert (2003); Marks and Niewiadomska-Szynkiewicz (2011)).

The radio signal propagation modeling

Propagation models are generally focused on predicting the average received signal strength at a given distance from the transmitter, as well as the variability of the signal strength in close spatial proximity to a particular location. Propagation models that predict the mean signal strength for an arbitrary transmitter-receiver separation distance are useful in estimating the radio coverage area of a transmitter and are called *large-scale* propagation models, since they characterize signal strength over large distances (hundreds or thousands of meters). On the other hand, propagation models that characterize the rapid fluctuations of the received signal strength over very short travel distances or short time durations are called *small-scale* models (see Rappaport (2002)).

In this paper we do not concentrate on small fluctuations of the signal strength in time. Hence the large-scale model is used further. Rappaport (2002) and many other authors claim that both theoretical and measurement based propagation models indicate that average received signal power decreases logarithmically with distance, whether in outdoor or indoor radio channels. The mean large-scale path loss can be expressed as a function of distance:

$$PL(d)[dB] = PL(d_0)[dB] + 10n \log\left(\frac{d}{d_0}\right), \quad (1)$$

where d is the transmitter-receiver distance, d_0 is a reference distance (for IEEE 802.15.4 radio typically the value of d_0 is taken to be 1 m) and n is the path loss exponent (rate at which signal decays). The value of n depends on the specific propagation environment and should be obtained through curve fitting of empirical data. Many authors, including Gibson (1999), indicate an empirical experiment as the best way to select an appropriate path loss for the reference distance d_0 .

The received signal strength P^r at a distance d is:

$$P^r(d)[dBm] = P^t[dBm] - PL(d)[dB], \quad (2)$$

where P^t denotes the power of transmitter.

Data collection

The data collection procedure was done by Jarosław Śmietanka – Warsaw University of Technology student during his thesis preparation (see Śmietanka (2015)). All series of experiments were done outside, in the field 10m x 10m. The MicaZ nodes were placed on wooden sticks 1.8 m tall. Two types of experiments were carried out. The first series was dedicated to measure RSSI values characterizing signal propagation between one mobile node and each of stationary nodes. As a result signal characteristics for three pairs of nodes were collected – they are described in *Training stage: RSSI-distance relationship identification* subsection. The second series of experiments was done assuming scenario with one mobile node moving around the whole test area and exchanging messages with three stationary nodes. This series of experiment is described in subsection *Testing stage: Tests on a full grid*. Figure 1 demonstrates the testbed configuration.

Training stage: RSSI-distance relationship identification

As it was written in previous subsection, the first experiment was dedicated to measuring RSSI values characterizing



Fig. 1. Testbed configuration for RSSI data acquisition.

signal propagation between one mobile node and each of stationary nodes. The aim of this experiment was identification of relationship between distance separating nodes and values of Received Signal Strength. At the beginning separation distance between nodes was equal 1 meter. Later it was extended one by one by 1 meter up to final separation distance equal 10 meters. The experiments confirmed the high variability of RSSI signal – what is illustrated in Figure 2. As it can be seen for mobile node and *stationary node #1* the registered values are not monotonous. For example the signal strength equal -75 dBm was observed both for distance equal 5 and 7 meters, while for 6 meters the signal was stronger (-73 dBm). Of course the presented values represents mean values for a series of 20-30 radio signal measurements.

Although the fluctuations of the signal strength make localization much more difficult, we decided to build the RSSI-distance relationship model and check how this fluctuations impose localization process. Similar results were observed by us earlier – Marks et al. (2014) and they didn't prevent us from obtaining high localization accuracy. The RSSI-distance modelling was done using OLS scheme (see Marks and Niewiadomska-Szynkiewicz (2011)).

Using (1) and (2) we can estimate the average distance between nodes i and j as a function of received signal strength P_{ij}^r :

$$\tilde{d}_{ij} = d_0 \cdot 10^{\frac{P^t - PL(d_0)}{10n}} \cdot 10^{-\frac{1}{10n} P_{ij}^r}, \quad (3)$$

where d_0 denotes the reference distance, $PL(d_0)$ the path loss at the reference distance, n the path loss exponent and P^t output power of the transmitter. It should be pointed that the goal of the calibration procedure is only to predict a value of the distance d_{ij} for known value of P_{ij}^r , not to find the exact value of the parameters $n, P^t, d_0, PL(d_0)$. Hence, we can simplify the equation (3) introducing parameters η and θ :

$$\tilde{d}_{ij} = \eta \cdot 10^{\theta \cdot P_{ij}^r}, \quad (4)$$

where $\eta = d_0 \cdot 10^{\frac{P^t - PL(d_0)}{10n}}$ and $\theta = -\frac{1}{10n}$. It seems to be reasonable to fit the RSSI-distance curve based on two parameters not four.

It is obvious that this average distance differs vastly from the true physical distance between selected nodes, but there is no chance to fit the curve describing signal propagation to all

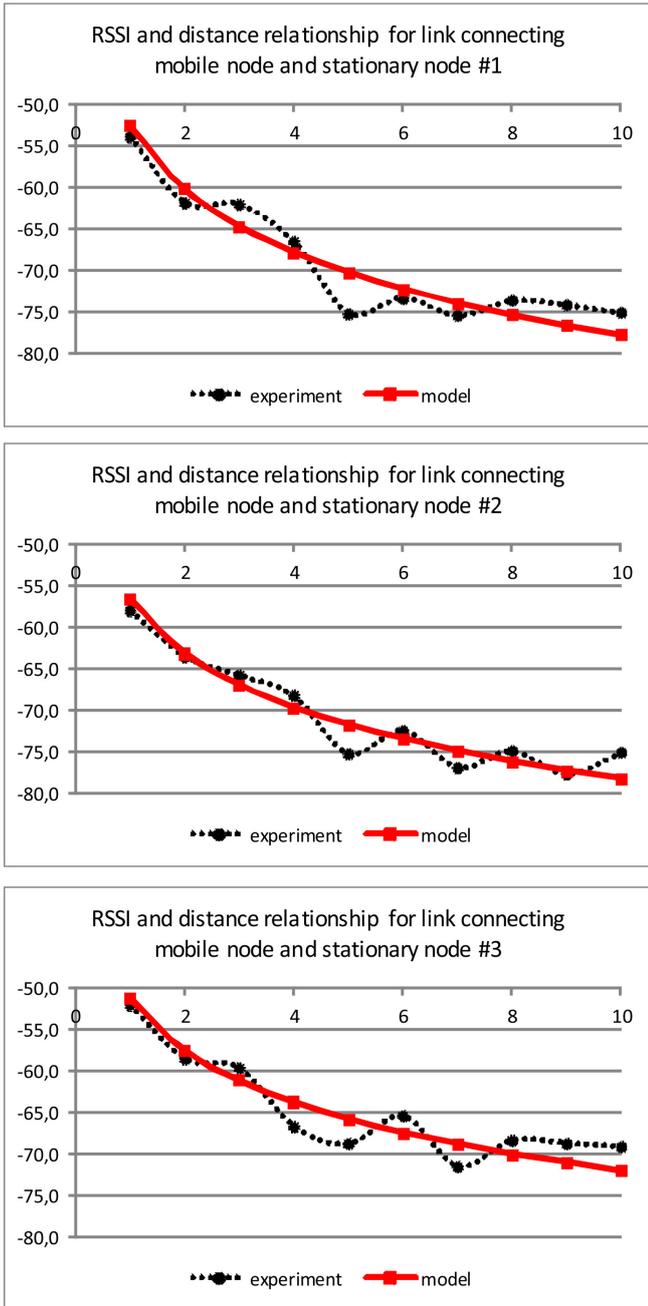


Fig. 2. Distance-RSSI relationship for Mobile node and particular anchors.

samples. An ordinary least square (OLS) method can be used to calculate values of parameters η and θ that minimize the error between the true physical and estimated distances:

$$\min_{\eta_{ols}, \theta_{ols}} \sum_{(P_{ij}^r, d_{ij}) \in \Psi} \left(\eta_{ols} \cdot 10^{\theta_{ols} \cdot P_{ij}^r} - d_{ij} \right)^2. \quad (5)$$

Obtained RSSI-distance curves are illustrated in Figure 2 by red lines.

Testing stage: Tests on a full grid

Using nomenclature from Machine Learning field the aim of training stage is building models which are later validated in testing stage. In the considered experiment the situation is very similar. The propagation models achieved as a solution of optimization task (5) creates basis to estimate distances in more realistic scenario. The estimation is done separately for each pair (mobile – stationary node). In this scenario, as it was written earlier, mobile node moves around the whole test area and exchanges messages with three stationary nodes. To be precise, by moving around, we understand a set of measurements which are done on a grid 10m x 10m with cell equal 1m x 1m. As a result at the end of this stage there are 121 locations with known distances between them and every stationary node. The experiment configuration is very similar to the one described by Bulusu et al. (2000).

LOCALIZATION TASK FORMULATION

For the purpose of simulation the robot-assisted localization we assume that the location of robot is known in each point of its trajectory and locations of three stationary nodes are unknown. The aim of localization process is determining the positions of stationary nodes with the minimal displacement. Without any information about relationship between stationary nodes the localization of each node can be treated independently. Hence, the average localization error can serve as the localization quality indicator.

Let's define the robot trajectory (path) A as a set of K points:

$$A = a_1, a_2, \dots, a_K, \quad (6)$$

where the distance between two neighbouring points is equal one meter:

$$\bigvee_{k=2, \dots, K} \|a_k - a_{k-1}\| = 1. \quad (7)$$

Therefore the localization task can be expressed as:

$$\min_{\hat{x}} J = \frac{1}{M} \sum_{i=1}^M \sum_{k=1}^K (\|a_k - \hat{x}_i\|_2 - \tilde{d}_{k,i})^2 \quad (8)$$

where M is the number of stationary nodes, \hat{x}_i denotes estimated positions of node i and $\tilde{d}_{k,i}$ distance between pairs of nodes (k, i) .

LOCALIZATION SCHEME

Classic Approach – Multilateration

The most intuitive approach to solve the task defined in (8) is to use collaborative multilateration algorithm described by Savvides et al. (2001). The idea of algorithm is very simple and it can be expressed as a minimization of differences between measured distances and distances resulting from estimated nodes locations:

$$\bigvee_{i=1, \dots, M} \min_{\hat{x}} \sum_{k=1}^K (\|a_k - \hat{x}_i\|_2 - \tilde{d}_{k,i})^2 \quad (9)$$

Real-data filtering

Multilateration is a widely used and popular algorithm in WSN localization, however its accuracy is not impressive for noisy measurements. Unfortunately in case of localization problem defined in previous section there is no place for algorithms utilizing the information about all connections in the network, which are much more precise, as the one described by Marks et al. (2014).

Therefore the only factor influencing localization accuracy which can be improved is the measurements quality. To be precise, not exactly measurements quality as we cannot change the data which are already obtained from real-life deployment, but the measurements can be filtrated to form a smooth curve. This is the consequence of condition (7) which limits possible changes in distances between stationary node x_i and two point on the path a_k, a_{k+1} to 1 meter. Only if the mobile node is going straight in direction of node s_i the distance is equal 1 meter. The general condition can be express as:

$$\left| \|d_{k,i} - d_{k+1,i}\|_2 \right| \leq 1 \quad (10)$$

In the proposed localization scheme the condition (10) is realized by application of two smoothing functions. The aim of the first one (**peak-filter**) is limitation local peaks in distances for subsequent a_k points. The second function applied after realization of (**peak-filter**) is (**local-change-filter**) which limits too big differences in subsequent distances values.

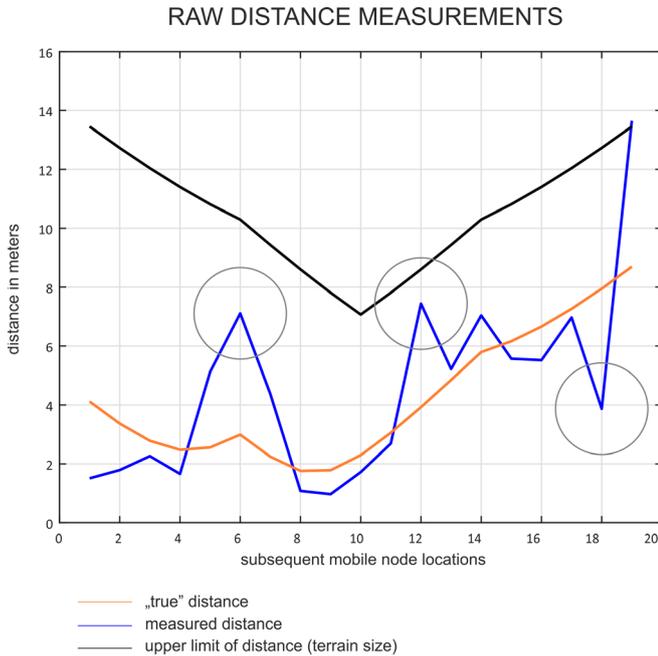


Fig. 3. Raw distance measurements obtained for subsequent mobile nodes positions.

To illustrate what is the distance filters purpose let's analyze the raw distance measurements obtained for subsequent mobile nodes positions and stationary node #1 – Figure 3. The grey circles denotes the highest peaks. The **peak-filter** identifies all peaks in the curve and limits them starting from

the most significant one. By significance we mean the absolute difference between values of distances for current, previous and next mobile node positions. How strict is the **peak-filter** function depends on the tolerance parameter which determines which absolute differences of distances are acceptable and which values should be corrected. Additionally **peak-filter** function utilize information about upper limit of distance – which is determined by terrain size. The node which is localized must be inside the searching field. The Figure 4 presents how the distance curve looks like after applying **peak-filter** – in upper part and after applying both filters in lower one.

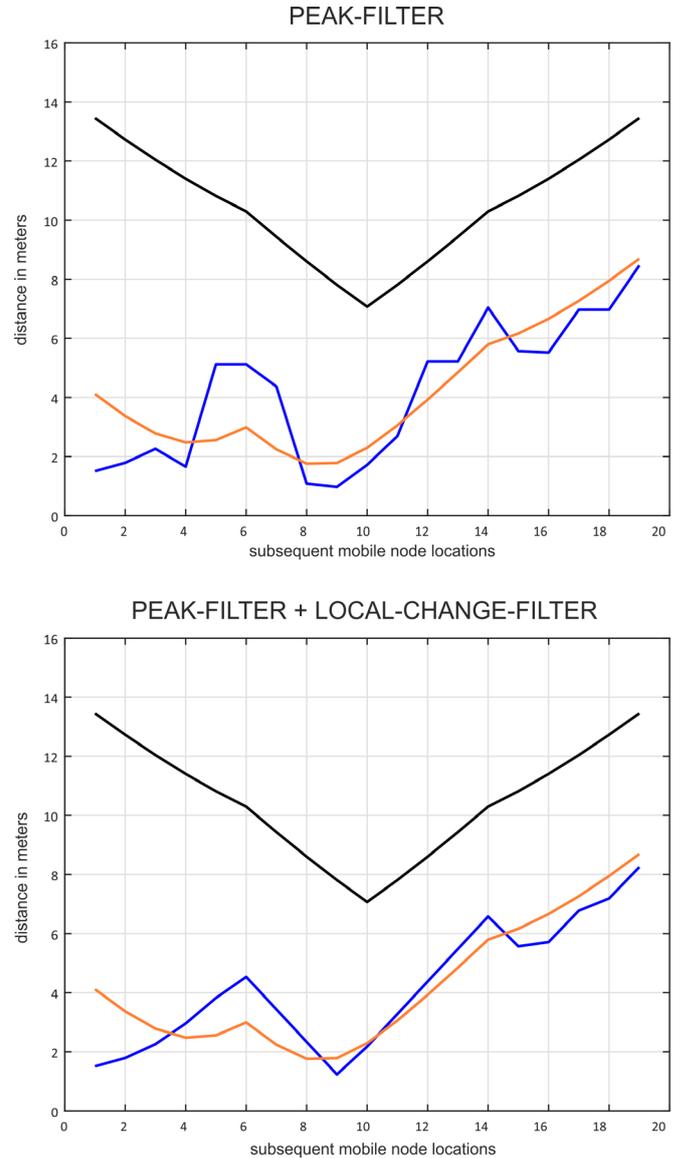


Fig. 4. Filtrated distance measurements obtained for subsequent mobile nodes positions.

The **local-change-filter** is used to determine and reduce changes in distances for subsequent mobile nodes positions occurring in monotonous function range. Similarly to **peak-filter** also **local-change-filter** behaviour depends on the tolerance parameter which determines what absolute differences of distances are acceptable and what values should be corrected.

NUMERICAL RESULTS

For the purpose of localization schemes evaluation the path composed of 19 locations was selected. The considered mobile node path is presented in figure 5. True locations of three stationary nodes are marked with blue (node #1), red (node #2) and green (node #3) triangles. The mobile node started in the bottom-left corner and moved to upper-right one. First attempt to localize stationary nodes was taken after visiting first four locations. Later the process was repeated until achieving the last location by mobile node. The final solution – after visiting all locations is presented in Figure 6. Figure 6a shows the locations obtained using raw distance measurements, while the Figures 6b and 6c illustrate locations found after localization method run on filtrated data – respectively after peak-filter and peak-filter + local-change-filter application. The shorter are lines connecting true locations (marked with triangle) and estimated locations (marked with squares) the better are location estimates. The quality of location estimates for different distance measurements (raw and filtrated) is presented in Table I. As it can be observed application of filtering process improves the localization quality, however for particular node the location error can be higher in comparison to RAW distance measurements – node #2. In general application of smoothing functions resulted in mean localization error reduction from 2.00 to 1.03 meter.

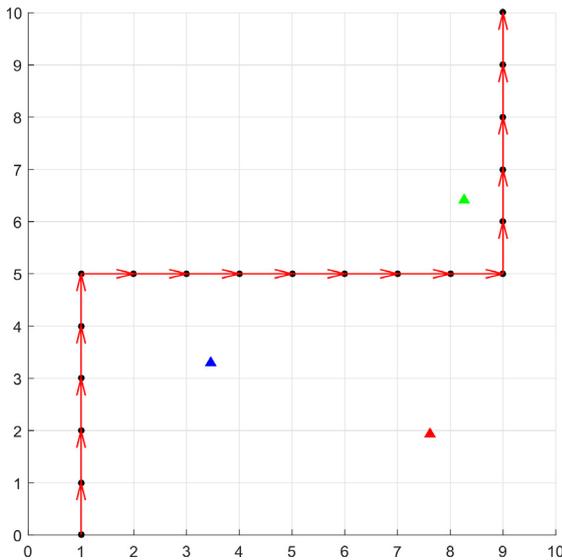


Fig. 5. Mobile node test path.

TABLE I. LOCALIZATION ERRORS FOR STATIONARY NODES 1-3 AND DIFFERENT DISTANCE MEASUREMENTS.

Localization Error	RAW distance measurements	Peak-filter	Peak-filter Intel + local-change-filter
Node #1	2.39	0.90	0.55
Node #2	1.90	2.09	2.38
Node #3	1.70	0.70	0.17
Mean	2.00	1.23	1.03

As it was mentioned earlier the localization process was repeated along the mobile node path. The values of localization errors as a function of travelled path length are presented in figure 7. The general rule is that the more measurements

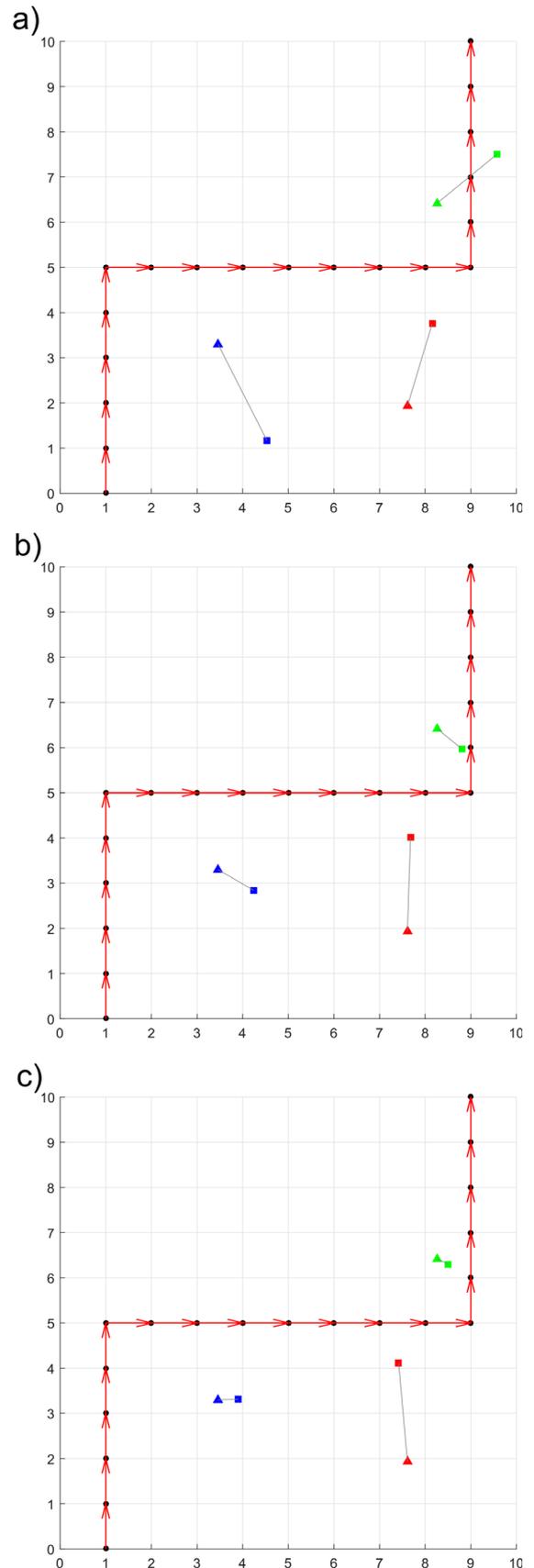


Fig. 6. Distance-RSSI relationship for Mobile node and particular anchors.

is available the more precise locations should be computed. However in some conditions the quality of localization can decrease after adding additional distance measurements which are disrupted by high measurement errors. Such situation can be observed for node #2 (red lines). The best position estimation was obtained after visiting 16 points along the mobile node path, not after 19 points. This is the result of significant underestimation of distances for the last 4 measurements – see row two in figure 8.

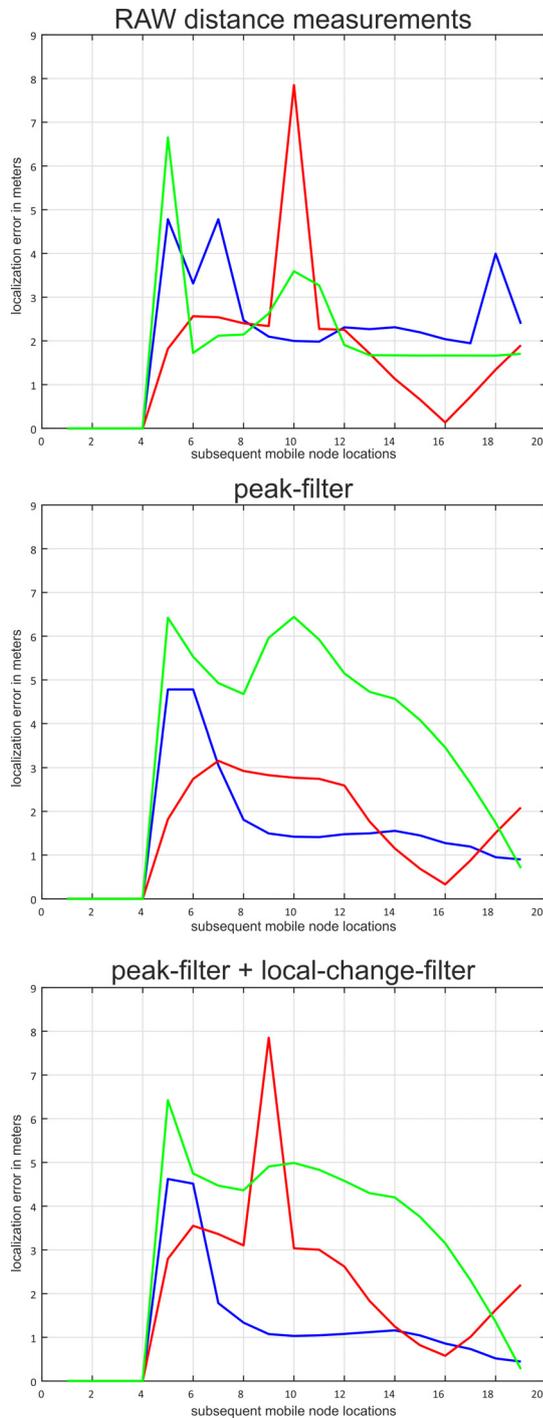


Fig. 7. Localization errors for different path length.

CONCLUSIONS

The aim of this study was simulation of robot-assisted localization in Wireless Sensor Network (WSN). For the purpose of simulation real-life data were collected and used to estimate inter-node distances exploited in further simulations. Such strategy guarantees that localization algorithms were validated in conditions not far from real-life deployments. In the same time it was possible to test localization methods for different mobile-robot paths without necessity of hardware paths realization. The main originality in our approach is incorporation of specialized smoothing filters. The proposed smoothing filters correcting distances measurements allowed on almost 50% localization error reduction in comparison to RAW distances measurements.

In the future we plan to prepare an extensive set of tests with different mobile node paths. These experiments should let us to use collected data to propose an optimal routes for mobile node, which allow us to minimize localization errors for stationary nodes.

ACKNOWLEDGMENT

We would like to thank MSc Jarosław Śmietanka for providing authors an access to data collected during his experiments with outdoor WSN deployments.

REFERENCES

- Benkic, K., Malajner, M., Planinsic, P. and Cucej, Z. (2008), Using rssi value for distance estimation in wireless sensor networks based on zigbee, *Systems, Signals and Image Processing*, 2008. IWSSIP 2008. 15th International Conference on, pp. 303–306.
- Bulusu, N., Heidemann, J. and Estrin, D. (2000), Gps-less low-cost outdoor localization for very small devices, *IEEE Personal Communications* 7(5), 28–34.
- Dewberry, B. and Petroff, A. (2015), Precision navigation with ad-hoc autosurvey using ultra wideband two way ranging networks, *Proceedings of IEEE Conference: 12th Workshop on Positioning, Navigation, and Communications (WPNC 2015)*, Dresden, Germany.
- Gibson, J. (1999), *The mobile communications handbook*, Electrical Engineering Handbook Series, second edition edn, CRC Press.
- Lazik, P., Rajagopal, N., Shih, O., Sinopoli, B. and Rowe, A. (2015), Alps: A bluetooth and ultrasound platform for mapping and localization, *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, SenSys '15*, ACM, New York, NY, USA, pp. 73–84.
- Liu, H., Darabi, H., Banerjee, P. and Liu, J. (2007), Survey of wireless indoor positioning techniques and systems, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 37(6), 1067–1080.
- Marks, M. and Niewiadomska-Szynkiewicz, E. (2011), Self-adaptive localization using signal strength measurements, *SENSORCOMM 2011, The Fifth International Conference on Sensor Technologies and Applications*, pp. 73–78.
- Marks, M., Niewiadomska-Szynkiewicz, E. and Kolodziej, J. (2014), High performance wireless sensor network localisation system, *Int. J. Ad Hoc Ubiquitous Comput.* 17(2/3), 122–133. <http://dx.doi.org/10.1504/IJAHUC.2014.065776>

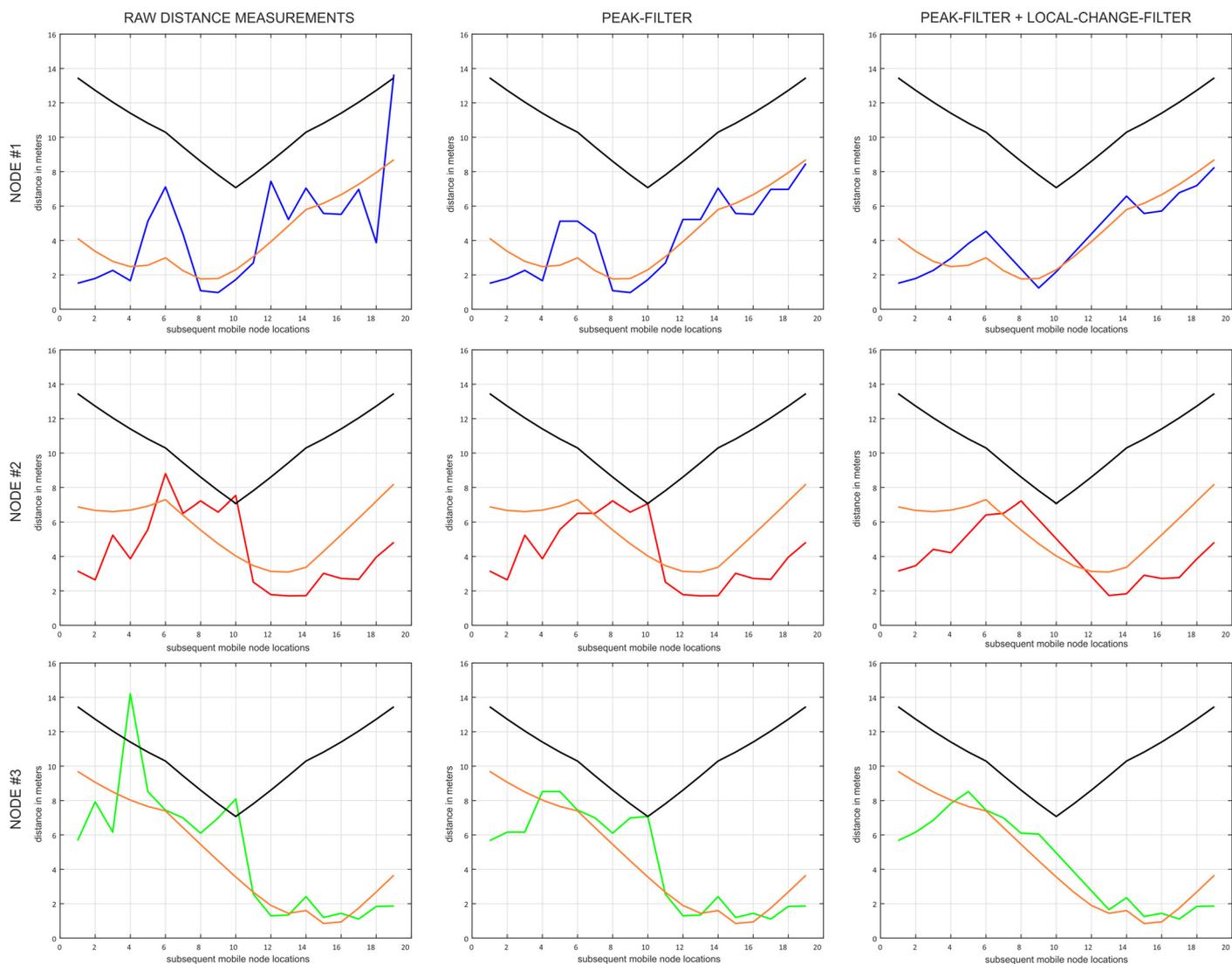


Fig. 8. Distance measurements obtained for subsequent mobile nodes positions. Each row represents distances for one stationary node. Different types of measurement – RAW and filtrated are presented in subsequent columns.

Ramadurai, V. and Sichitiu, M. L. (2003), Localization in wireless sensor networks: A probabilistic approach, Proceedings of International Conference on Wireless Networks (ICWN 2003), Las Vegas, pp. 300–305.

Rappaport, T. (2002), *Wireless communications: principles and practice*, Communications Engineering and Emerging Technologies Series, second edition edn, Prentice Hall.

Sánchez, C., Ceriani, S., Taddei, P., Wolfart, E., Rodriguez, A. L. and Sequeira, V. (2015), *Steam: Sensor tracking and mapping*, Seattle, WA, USA.

Savvides, A., Han, C.-C. and Strivastava, M. B. (2001), Dynamic fine-grained localization in ad-hoc networks of sensors, Proceedings of the 7th Annual International Conference on Mobile Computing and Networking, MobiCom '01, ACM, New York, NY, USA, pp. 166–179.

Śmietanka, J. W. (2015), Nodes localization in heterogenous ad hoc network (in polish – lokalizacja węzłów heterogenicznej sieci ad-hoc), Master thesis, Warsaw University of Technology, Warsaw.

AUTHOR BIOGRAPHIES



MICHAŁ MARKS received M.Sc. (2007) in computer science and Ph.D. (2015) in automation and robotics from the Warsaw University of Technology. Since 2007 with Research and Academic Computer Network (NASK). The author and co-author of over 30 journal and conference papers. 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.

PERFORMANCE EVALUATION OF SOA IN CLOUDS

Ashraf M. Abusharekh
Dalhousie University
Halifax, Nova Scotia
Canada
ashraf@dal.ca

Alexander H. Levis
George Mason University
Fairfax, Virginia
USA
alevis@gmu.edu

KEYWORDS

Service Oriented Architecture, Architecture Federation, Multi-formalism Modeling, Measures of Performance, Measures of Effectiveness, Quality of Service.

ABSTRACT

An approach for constructing an Enterprise Service Bus based Service Oriented Architecture for individual clouds and then considering workflows where services from multiple clouds are used is formulated and a systematic methodology for performance evaluation of the architecture is presented. The participation in this environment is achieved by allowing the SOA to dynamically federate with services through Community of Interest registries, i.e., different clouds, and by utilizing these services to share enterprise-level information. The performance prediction methodology is based on multi-formalism modeling.

INTRODUCTION

Service orientation enables enterprise interoperability and resource re-use. However, as clouds proliferate, each with its own core and application-specific services, the problem of invoking services and creating workflows that use services from two or more clouds needs to be addressed. Consequently, a methodology to evaluate and predict the logical, behavioral, and performance characteristics of such a federated Service Oriented Architecture (SOA) becomes a necessity. Such a methodology would synthesize an executable model capable of capturing the complexity of a SOA *federation* deployed on different clouds.

The system architect or designer needs to analyze the *dynamic* behavior of the proposed workflow, identify logical and behavioral errors not easily seen in the static documentation of the architecture, and demonstrate the capabilities that the architecture enables and how well they could be used. What makes the problem challenging is that the system's behavior and performance don't only depend on the system's services but also on (a) the infrastructure services that enable loose coupling, (b) services implemented by other systems, i.e., residing in other clouds, and (c) the underlying network supporting the different cloud architectures. Sustaining acceptable end-to-end performance in such a dynamic environment becomes a challenging problem.

This paper presents a methodology for performance evaluation and prediction of an Enterprise Service Bus (ESB) enabled SOA federation. The methodology involves the development and implementation of a multi-formalism based executable model that is capable of

capturing and predicting the dynamic behavioral and performance aspects of a SOA federation. The executable model aids the system architect in debugging and evaluating the architecture, and helps verify that the proposed architecture will satisfy requirements and how well it will do so.

RELATED WORK

Liu et al. (2007) presented a performance modeling and benchmarking approach that facilitates estimating performance characteristics of the Enterprise Service Bus and analyzing the performance relationship between the ESB and its composite applications. To simplify the performance analysis, their work focuses only on the performance of ESB routing and transformation; this model does not incorporate the orchestration service, or modeling of the technological network.

Sloane et al. (2007) presented a hybrid approach to modeling SOA systems of systems in which two separate models are developed, a Colored Petri Net (CPN) model used to capture internal protocols, communications and resource consumption and a discrete event simulation model called MESA (Modeling Environment for SOA Analysis) to capture the interactions between nodes in a SOA environment. Although called a hybrid approach, the CPN and the MESA models are completely isolated. This approach is hard to generalize to capture different behaviors, and doesn't fully capture the effect of the network on the behavior and performance of a SOA. Finally, the approach does not capture business processes, the very driver of employing cloud-based service orientation in the first place.

Shin and Levis (2003) use CPN and network simulator models to gain insight into the behavior and performance characteristics of architectures. Their approach has two separate executable models, the functional executable model is a CPN and the physical (communication) model is a queuing net modeled using the ns-2 (2008) network simulator. The simulation models run offline, i.e., no message exchange between the two executable models during run-time.

Abusharekh, et al. (2009) presented an approach to evaluating the end-to-end response time of business processes deployed in an ESB-enabled SOA environment. The discrete event simulator OMNeT++ (2009) provided the behavior environment in which SOA-based performance was evaluated. An abstract ESB model which supports business process orchestration, routing, and reliable messaging was introduced. The model was capable of specifying the SOA supporting network to the needed level of detail. This approach assumed that

an architecture had been designed and that its description consisted of a set of static views describing the logic and behavior of the business services and business processes and a physical view.

Ghasemi, et al. (2014) transformed UML activity diagrams to generalized stochastic Petri Nets (GSPN); although their approach is targeting SOA performance evaluation, it does not incorporate the complexities of SOA ESB, or capture the overhead associated with the cloud and network services. Duan (2015) focused on the critical role of network communications in cloud computing and its effect on end-to-end performance of cloud service provisioning. He presented an approach for characterizing the service capabilities of a composite network-computer system using network calculus. The approach targets the cloud platform and does not fully capture the architecture to be deployed to the cloud infrastructure. Bocciarelli, et al. (2015) introduced a model-driven approach to generate HLA-based simulation from SysML specifications of autonomous systems.

The present work extends the work done by Abusharekh, et al. (2009) by employing multi-formalism modeling in which CPN and network models interoperate during execution. Our approach goes further in allowing the architect to capture the specifics of the communication network at any level of detail through the network simulator. Furthermore, our approach allows the architect to capture the SOA components and the underlining business processes and their interaction with the technological network.

THE DESIGN PHASE

In a SOA federation, SOAs co-exist in different clouds. Each SOA has established producer-consumer relationships, such that the right rules and policies (trust, governance, security, etc.) apply throughout its environment. This allows for autonomy of individual SOAs but requires implementing federation-wide rules and policies to regulate and govern the federation. (Erl, 2004; Goodner and Nadalin, 2007)

The concept of Community of Interest (COI) is used to enable dynamic federation with pre-defined or un-anticipated systems. A Community of Interest is a problem-solving entity that utilizes services across different SOAs to implement the workflow that addresses its problem. In order to simplify and speed-up the discovery of services, COIs will not only define common vocabularies, taxonomies, data standards, interchange agreements, and specifications among COI members, but also will define service descriptions relevant to the communities and will host a repository of current implementations of those services. Each COI will have its federation repository.

To further clarify the SOA federation construct and how the notion of COI enables and supports such an environment, an example is depicted in Fig. 1. Several problems with this approach need to be resolved such as the location of a COI repository that should be negotiated and agreed upon among participating parties. The

Cloud environment is assumed to host the COIs and the Cloud registry is the central federation repository/registry that publishes COI information. When a service failure occurs candidate services are examined at other SOAs to locate an alternative service implementation.

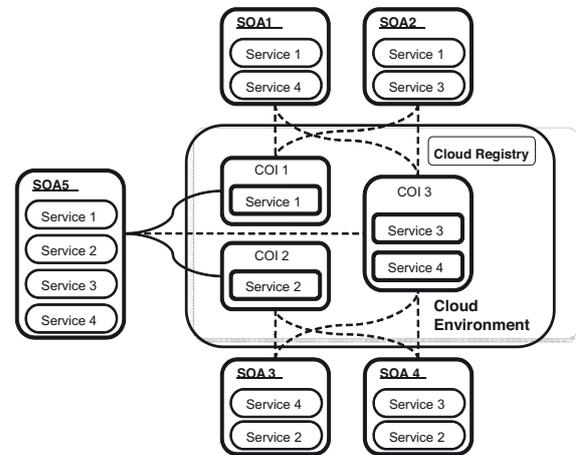


Fig. 1 Cloud Environment and SOA Federation

The objective of the approach is to construct an event-driven SOA capable of participating in the Cloud environment by consuming existing services and/or to populate it with new ones that can be consumed by anticipated and un-anticipated users. This SOA infrastructure is based on an Enterprise Service Bus (ESB). The services and processes are published through their Communities of Interest (COI) as services for other SOAs and COIs to re-use.

Two additional sources of information are needed to be able to insert a new SOA in the Cloud environment:

- (1) Information about existing COIs and the services they expose in order to be able to consume services and to publish new ones. Full understanding of the other COIs policies and rules, their data formats, and services descriptions is needed to successfully federate with them; the new SOA will need to abide by them. The Cloud Registry Service will be the main source of such information. The designer needs to understand the services exposed by other SOAs in order to make a decision whether or not they will fulfill the functional and non-functional requirements of the new federate.
- (2) Access to existing Enterprise Services currently available in the Cloud environment. These enterprise services will allow for trustworthy enterprise-level data, information, and services sharing.

The products of the Design phase are then used in the Analysis and Evaluation phase to construct an executable model. Structural and behavioral models of a system architecture (in this case of a workflow) are static representations of an unprecedented, complex, and dynamic system. These models are capable of describing the behavior of the architecture only in a limited way. Consequently, there is need for evaluation techniques that go beyond static diagrams. An executable model of the

workflow design enables the designer to analyze its dynamic behavior, identify logical and behavioral errors not easily seen in the static descriptions, and demonstrate to the user the capabilities that the workflow enables.

THE ANALYSIS AND EVALUATION PHASE

A key concept in the Analysis and Evaluation phase is that all elements of the executable model must be traceable to elements in the architecture description of the workflow; any corrections or changes introduced in the executable model should be reflected back in the architecture data.

In order to predict the dynamic behavior and performance of a process or application composed of multiple services, behavior and performance characteristics of participating services and the supporting network must be captured. To accomplish this, the executable model makes use of two model formalisms, a discrete event system model expressed as a Colored Petri Net and a network model expressed as queueing network.

Inputs to the Analysis and Evaluation phase are the structural and behavioral models that describe the functional, the services, and the systems views of the architecture. Outputs of this phase are either changes to the architecture, or an architecture ready for deployment along with its Measures of Performance (MOPs) and Measures of Effectiveness (MOEs).

Stage 1: Behavioral and logical evaluation. Verification of the logical and behavioral aspects of the architecture must be done before performance evaluation. An executable model of the business processes and services is built using CPNTools (2016). The inputs to this stage are the data in the functional architecture models, and the outputs are the services definitions, the services implementations as CPN Models, and the business processes as XML files, all of which are fed to Stage 2.

Step 1: Synthesizing CPN Executable Model

The structure of the CPN model includes organizational nodes, the services under their organizational boundaries, and the business processes they own. Although the CPN model is for the functional viewpoint of the workflow design, Services and Systems models are also needed to define and construct services definitions, interfaces and implementations.

Two types of services are modeled in CPN, singleton services and composite ones. A singleton service has one input place representing service requests and one output place representing service response. A composite service has two additional places, an output place representing a process request and an input place representing a process response. For simplicity, each service is assumed to have one and only one activity (function) and the current composite service model allows for one process to be requested.

Step 2: Evaluation using the CPN Executable Model

At this step, scenarios need to be defined to evaluate the logical and behavioral aspects of the design and any logical or behavioral errors will be captured and fixed and

be reflected directly to the architecture models to maintain traceability. The state space analysis tool embedded in CPNTools is used to generate and examine the model's state space to detect errors or unwanted behavior. After the behavioral and logical analyses are done, processing delays of the services are added to the CPN model (Timed CPN) and the performance of the processes is analyzed. This performance reflects processing delays of services only; additional overhead due to SOA infrastructure services and the underlying network infrastructure is not captured here. If the performance of the CPN model does not meet the requirements of the architecture, the designer must make changes to the architecture to improve the performance. This CPN model performance serves as the best case (baseline) performance of the design; adding the SOA and network infrastructure will degrade performance.

Stage 2: Performance prediction and evaluation. The inputs to this stage are the outputs of Stage 1 and the Services viewpoint models; the outputs are changes to the design. The tools suite used at this phase is the C2 Wind Tunnel. (Balogh et al., 2008) This is a High Level Architecture (HLA, 22000) simulation environment that integrates various simulation platforms, including CPNTools and OMNeT++. The C2 Wind Tunnel integration is done on three levels: the API level, the interactions level, and the model semantics level. API level integration provides basic services such as message passing and shared object management, while the interaction level integration addresses the issues of time synchronization and coordination. Semantic integration is more subtle and depends upon the goals and the context of the overall simulation environment. At the API and interactions level the C2 Wind Tunnel uses Portico (2009) an open-source, cross-platform HLA RTI implementation. At the level of model semantics it uses the meta-programmable Generic Modeling Environment (GME, 2009) engine to integrate the operational semantics of multiple simulation platforms, to manage the configuration and deployment of scenarios in the simulation environment, and to generate the necessary interface code for each integrated simulation platform. GME is used to generate configuration files and HLA interactions and glue code for the C2 Wind Tunnel to host the two interoperating models that make up the multi-formalism based executable model. Configuration is accomplished through meta-models that formally specify the modeling paradigm of the application domain. A meta-model is used to define all the syntactic, semantic, and presentation information regarding the domain.

The goal is to compute the creation time (T_C) and end-to-end response time (T_R) of processes deployed on such an environment as defined in Abusharekh, et al. (2009). T_C and T_R depend on the behavior and performance of the ESB services and the business services contributing to the business process, the underlying network supporting the SOA and the cloud environment, and the request load of business processes deployed on the SOA at a given time. In order to capture the above factors and the related characteristics of the environment, five profiles

need to be created: (1) Network profile, (2) ESB profile, (3) Services Profile, (4) Processes Profile and (5) Scenario profile. Abusharekh et al. (2009) provide a full description of the profiles and their structure.

Step 3: Building the Multi-formalism based Executable Model

In the executable model services are modeled in CPN-Tools while the network infrastructure is modeled in OMNeT++. The C2 Wind Tunnel instance that was used hosts two types of federates, CPN federates and Network federates. The structure of the network model is shown in Fig. 2. MOM is the Message oriented Middleware module.

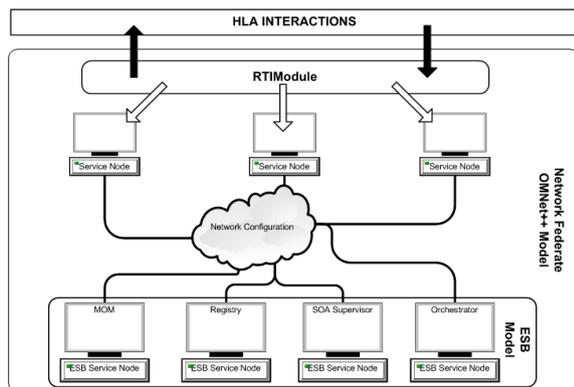


Fig. 2 OMNeT++ Model Structure

The OMNeT++ model implements the RTI module, a generic SOA ESB model, service nodes, and the network topology. The RTI module supports communication with the C2 Wind Tunnel through HLA interactions. The ESB model is the same as that introduced in Abusharekh et al. (2009) but enhanced by including a Service Application module managing interactions from/to the CPN model through the RTI Module.

Step 4: Evaluation Using the Multi-formalism based Executable Model

This step needs a scenario profile to define the request loads on different business processes and within different classes of a business process. The results of the execution are analyzed, after which the designer either is satisfied with the results or decides to make changes to one or more of the profiles in order to improve the overall performance.

The performance evaluation of the architecture is done using the System Effectiveness Analysis methodology (Levis, 1997). The MOP of interest are *Timeliness* and *Accuracy* of a federated SOA architecture. Accuracy is defined as the expected cost of a business process producing an outcome different from the desired one. The variables quantifying timeliness are business process response time, creation time, and throughput rate. Identifying variables quantifying accuracy depends on the specific mission and objectives of the workflow.

After the executable model has been configured, the Simulation panel of the C2 Wind Tunnel is used to generate all necessary files for the executable model to run.

A CASE STUDY

The case study results presented here are based on a hypothetical operational concept for a system called Airborne Theater Ballistic Missile (TBM) Interceptor System (ATIS). This case study is used here because a documented complete architecture exists. (Abusharekh et al. 2007)

In order to create a cloud based architecture capable of intercepting and destroying TBMs and capable of providing and consuming information and services to and from this particular cloud environment we need to: (a) define services and processes to be hosted by the ATIS (internal services); (b) re-use business services and/or processes implemented by other and residing in different clouds systems (external capabilities); and (c) publish relevant ATIS services to be used by other systems in this cloud environment. To re-use existing services and populate the cloud environment with new ones, ATIS must join relevant COIs that have their own SOA in different clouds. Two COIs are modeled: (a) the Ballistic Missile Response (BMR) COI: a collaborative group of cloud users who exchange information regarding ballistic missile response; (b) the Intelligence, Surveillance and Reconnaissance (ISR) COI: a collaborative group of cloud users who exchange information related to ISR.

The operational concept graphic of the architecture is shown in Fig. 3. It is assumed that Core Enterprise Services (CES) are available and accessible. The Systems and Services Interface Description is shown in Fig. 4 as a Unified Modeling Language Deployment diagram.

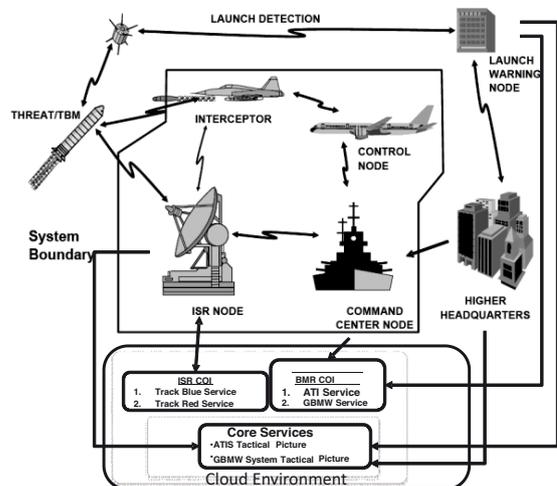


Fig. 3 Operational Concept Graphic for ATIS

A CPN executable model using CPNTools was created to evaluate the operational aspects of the design. This first model included the processes and services of the ATIS, but no ESB interaction was included. Once created, the executable model was used to check the logic and behavior of the services and their composition into processes. As errors were detected, fixes were made to the CPN model and reflected back to the architecture models. The performance of the ATIS capabilities was tested by converting the CPN model into a Timed CPN.

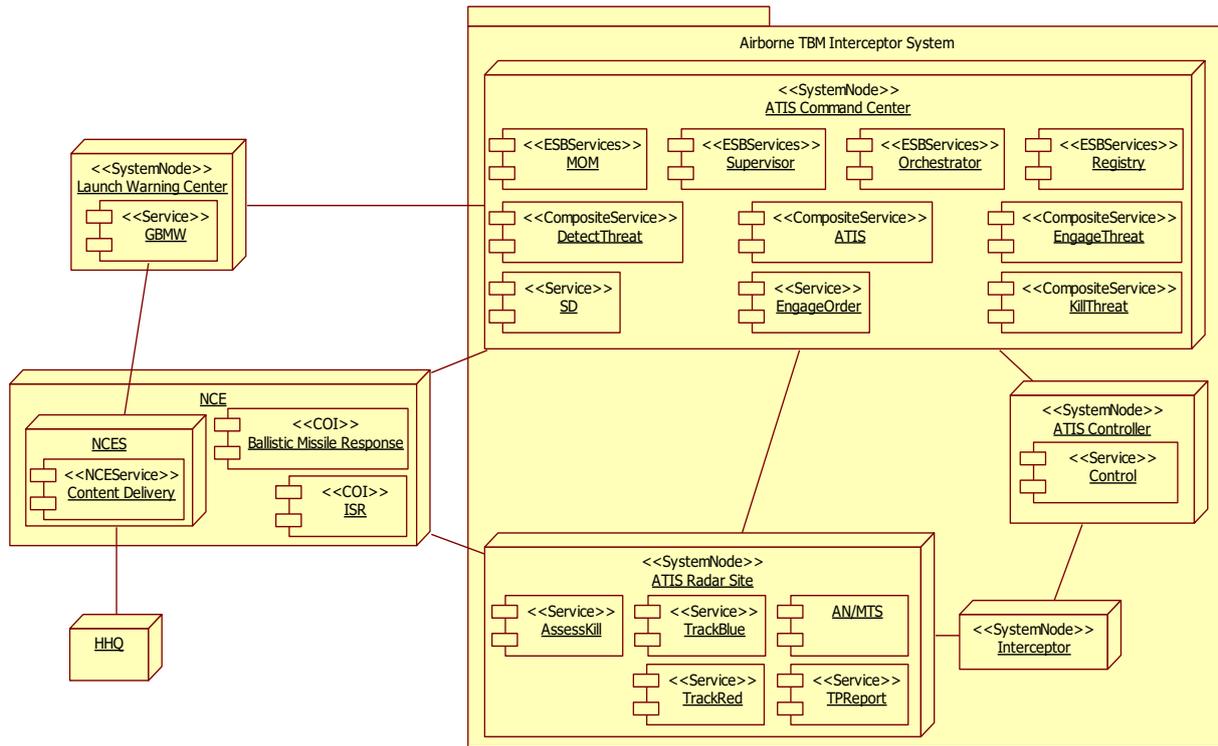


Fig. 4 Services & Systems Interface Description

The main questions to be addressed are: (1) Is the ATIS capable of intercepting in a timely manner incoming TBMs? (2) How many interceptors are required to handle various adversary capabilities while keeping the average response time and number of leakers within the requirements? Therefore, the parameters of interest are:

1. TBM inter-arrival time: a continuous parameter with values used in the experiments of 0, 25, 50, 75 and 100 seconds.
2. The number of ATIS interceptors: a discrete parameter with values 3, 4 and 5 interceptors.

The resulting parameter locus is shown in Fig. 5. It consists of the three vertical lines.

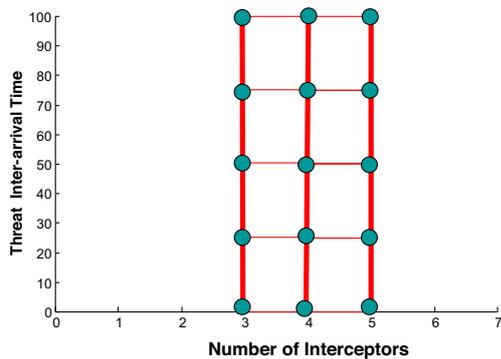


Fig. 5 The Parameter Locus

Three Measures of Performance were used to assess ATIS: (a) *Average response time*: the average time between ATIS detecting the threat and the threat being engaged by ATIS interceptors. The requirement is for ATIS to be able to destroy a TBM within 400 seconds of detecting it. (b) *Accuracy*: the number of leakers

(TBMs not destroyed after 400 seconds of being detected by ATIS). This must be no more than 2. (c) *Throughput rate*: the number of TBMs processed per unit time. There is no constraint so it is parameterized.

The following scenario was used for performance analysis. An adversary capable of launching multiple missiles exists and the ATIS and the Global Ballistic Missile Warning service provided by the BMR COI have been deployed. The summary of the results of the 15 simulation runs for the baseline case in which the SOA and network infrastructures are not present are in Table 1. They show that 3 interceptors can handle the 10 threats with a maximum of two leakers if the threats arrive at a rate slower than 1 in 25 seconds. These results represent the best performance the architecture could accomplish.

Table 1. Number of Leakers vs. Number of Interceptors

# of Interceptors	TBM Inter-arrival	Average Response Time	# of Leakers
3	0	347.1	4
	25	270.1	1
	50	180.6	0
	75	159.0	0
	100	159.0	0
4	0	283.9	2
	25	212.9	0
	50	159.0	0
	75	159.0	0
	100	159.0	0
5	0	245.5	0
	25	180.0	0
	50	159.0	0
	75	159.0	0
	100	159.0	0

The multi-formalism based executable model captures the communications systems and the SOA infrastructure and shows how they will interact to enable the composition of services in processes for successfully executing the mission.

To predict the performance of federated SOAs, the scenario was modified so that the first ATIS unit, ATIS_A, federates with another unit of ATIS, ATIS_B in order to overcome failures in services during an attack. The two ATIS units cover adjacent geographical regions and are both members of the ISR COI. The region under ATIS_A is under attack, and after 350 seconds, ATIS_A's ISR node fails. ATIS_A then federates with ATIS_B (not under attack) and re-uses its *TrackRed*, *TrackBlue*, *TPReport* and *AssessKill* Services published through the ISR COI. The goal is to show the effect of SOA federation on the performance of the architecture.

Figure 5 shows the average response time of the ATIS for three different message sizes (1KB, 500KB, and 1000KB with the latter containing images) and how it is affected by increasing the number of interceptors under the federated network infrastructure. With long TBM inter-arrival times, increasing the number of interceptors produces no significant gain in response time for large message sizes. The best achievable performance for different messages sizes is:

1. 1KB message sizes, average response time of 184.5 sec with no leakers.
2. 500KB message sizes, average response time of 238 sec with 4 leakers.
3. 1000KB message sizes, average response time of 290 sec with 10 leakers.

Sensitivity analysis of the number of leakers to the number of ATIS interceptors and message size, respectively, with different adversary launch capabilities, showed that, with large message sizes, increasing the number of interceptors will not decrease the number of leakers.

For each operating point in the parameter locus, a value for each Measure of Performance is computed using the executable model. The set of all such values is the Performance Locus. The Requirements locus is the set of admissible values of the two measures of performance: Average Response Time ≤ 400 ; Leakers ≤ 2 .

The ATIS Measures of Effectiveness are calculated by comparing the measures of performance against the requirements. (Levis, 1997) This is computed as the fraction of the Performance locus that intersects the Requirements locus.

ATIS Requirements and Performance loci for the single SOA without communication delays as a function of message size are shown in Fig. 7. Considering the performance requirements for the average response time and the number of leakers, the Measure of Effectiveness of the single SOA model can be calculated by considering the projection of the performance locus on the average response time against the number of leakers. The resulting value is 93%, i.e., the single SOA architecture without communication delays is 93% effective.

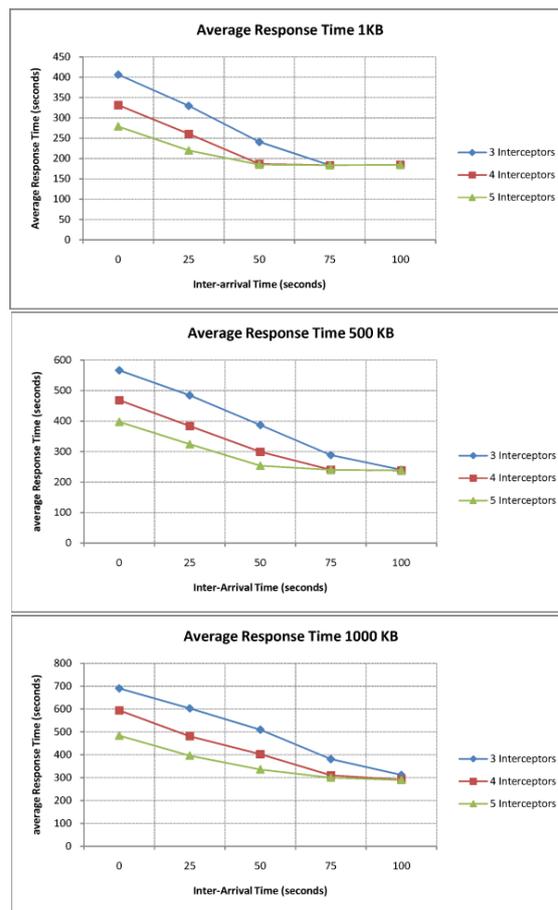


Fig. 6 Average Response Time vs. Inter-arrival time

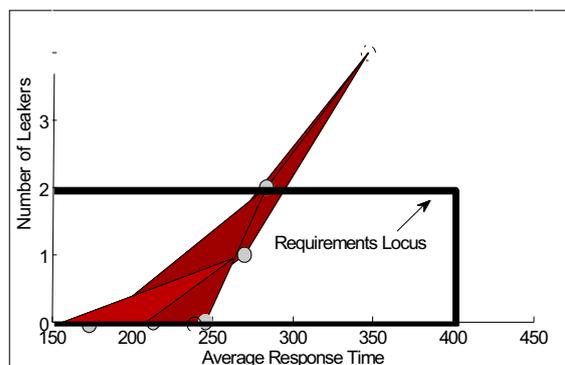


Fig. 7 Evaluation of Measure of Effectiveness

When the same analysis is done for the federated SOA architecture, the Measure of Effectiveness drops to 73% for 1KB message size and to 0% for the larger message sizes. This is a significant result because it shows that the infrastructure (computers and communications) needed to implement a federated SOA architecture exact a very substantial Quality of Service cost. For time sensitive operations, this cost may be unacceptable. For other types of operations, this type of cost should be considered when determining whether to use federated

SOAs, i.e., creating workflows that use services residing in different clouds, or not.

CONCLUSIONS

This paper introduced a *SOA federation framework* that allows dynamic federation between SOA instances that are implemented in different clouds. This design enables the provision of new services and the consumption of existing ones through dynamic federation.

The evaluation approach is based on synthesizing an executable model that is capable of capturing and predicting the dynamic behavioral and performance aspects of the Service Oriented Architecture. The executable model employs multi-formalism that makes use of two models, CPN models for the SOAs and the workflows and an OMNeT++ network model that captures the underlying technological infrastructure in some detail.

One of the main conclusions of this analysis is that using the multi-formalism based executable model gave more insight and understanding of the dynamics of a federated SOA design. While SOA offers many advantages such as information sharing, reuse, and interoperability it also has its drawbacks. More information sharing across the federated SOAs introduces more latency and can cause substantial degradation in the quality of service (QoS). For time-critical system such as ATIS, more information sharing is not sufficient to produce better actions due to time constraints. Services receive better information but too late to act on it.

The critical conclusion is that while a single SOA offers many performance and agility advantages, a federation of SOAs introduces Quality of Service issues that may indicate that this is not always the right solution. Thus it is necessary to analyze alternative architectures to determine the appropriate solution for a particular problem. The executable model presented is capable of verifying, evaluating, and demonstrating the capabilities of a SOA and how well it performs as a single instance and also as part of a federation. In addition, the executable model can be used as a test-bed to evaluate new algorithms and protocols to enhance the design.

With the emergence of cloud and containerization (PaaS/IaaS) technologies, deploying a federated SOA that satisfies the functional and non-functional requirements is becoming a challenge for the architect/designer. Future work includes exploring and evaluating the use of these technologies to achieve the targeted non-functional requirements of the architecture, such as timeliness and availability. COI concepts can be leveraged to allow dynamic evolution of the federated SOA at runtime to fulfill its mission or the COI mission, e.g. using service mirroring to improve performance and/or availability. These issues need to be explored and evaluated in a quantitative manner, before investing in implementing and deploying the architecture.

REFERENCES

Abusharekh, A., S. Kansal, A.K. Zaidi, and A.H. Levis, 2007. "Modeling Time in DODAF Compliant Executable Architectures," in *Conf. on Systems*

- Engineering Research*, Hoboken, NJ.
- Abusharekh, A.; L. E. Gloss; and A.H. Levis, 2009. "Evaluation of SOA-based Federated Architectures," *Systems Engineering*.
- Balogh, G. et al., 2008. "Rapid Synthesis of HLA-Based Heterogeneous Simulation: A Model-Based Integration Approach.," Unpublished manuscript.
- Bocciarelli, P., A. D'Ambrogio, A. Giglio, and E. Paglia. 2015. "A model-driven framework for distributed simulation of autonomous systems." In Proc. Symp. on Theory of Modeling & Simulation: DEVS Integrative M&S Symposium (DEVS '15), F. Barros, M. H. Wang, H. Prähofer, and X. Hu (Eds.). Soc. for Computer Simulation International, San Diego, CA.
- CPN Tools, 2016 [Online]. <http://cpntools.org/>
- Duan, Q. 2015. "Modeling and performance analysis for composite network-computer service provisioning in software-defined cloud environments." *Digital Communications and Networks*, (1)3.
- Erl, T., 2004. *Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services*, Prentice Hall.
- Ghasemi, A., A. Harounabadi and S. J. Mirabedini. 2014. "Performance Evaluation of Service-Oriented Architecture using Generalized Stochastic Petri Net." *Int. J. of Computer Applications* 89(6).
- GME, 2009. Generic Modeling Environment. [Online]. <http://www.isis.vanderbilt.edu/projects/gme/>
- Goodner, M. and A. Nadalin, 2007. "Web Services Federation Language (WS-Federation) Version 1.2," September 26.
- HLA, 2000 IEEE Standard for Modeling and Simulation (M&S) High Level Architecture.
- Levis, A. H. 1997. "Measuring The Effectiveness of C4I Architectures," in 1997 *Int. Symp. on Defense Information*, Seoul, Republic of Korea.
- Liu, Y.; I. Gorton, and L. Zhu, 2007. "Performance Prediction of Service-Oriented Applications based on an Enterprise Service Bus," in *Computer Software and Applications Conference*, vol. 1, Beijing, pp. 327-334.
- NS-2, 2016. The Network Simulator - ns-2. [Online]. <http://www.isi.edu/nsnam/ns/>
- OMNet++ 2016. [Online]. <http://www.omnetpp.org/>
- Portico, 2009. The Portico Project. [Online]. www.porticoproject.org/
- Shin, I. and A.H. Levis, 2003. "Performance prediction of networked information systems via Petri nets and queuing nets," *Systems Engineering*, vol. 6, no. 1, pp. 1 - 18.
- Sloane, E.; T. Way; V. Gehlot; R. Beck; J. Solderitch; and E. Dziembowski, 2007. "A Hybrid Approach to Modeling SOA Systems of Systems Using CPN and MESA/Extend," *Systems Conference*.
- Dr. Ashraf M. Abusharekh** is a senior research scientist at the Faculty of Computer Science at Dalhousie University, NS, Canada. **Dr. Alexander H. Levis** is University Professor of Electrical, Computer, and Systems Engineering at George Mason University, Fairfax, VA., USA.

THREE LAYERS NETWORK INFLUENCE ON CLOUD DATA CENTER PERFORMANCES

Marco Gribaudo
DEIB

Politecnico di Milano
via Ponzio 51
20133, Milano, Italy
marco.gribaudo@polimi.it

Mauro Iacono
DSP

Seconda Università degli Studi di Napoli
viale Ellittico 31
81100 Caserta, Italy
mauro.iacono@unina2.it

Daniele Manini
DI

Università degli Studi di Torino
corso Svizzera 185
10129, Torino, Italy
manini@di.unito.it

KEYWORDS

cloud networking; data center performances; Markovian agents; performance modeling; virtualization

ABSTRACT

The effects of networks on the performances of cloud architectures are a very significant issue in designing a data center. The efficiency of data transfers and the overall traffic management are a critical factor that constitutes a potential performance bottleneck, potentially limiting the number of computing nodes that can be installed more than their cost issues. In this paper we present a modeling approach, based on Markovian agents, that allows a performance analysis of network effects in high scale cloud architectures.

I. INTRODUCTION

Virtualization is a key technology in the field of cloud computing. The use of virtual machines (VM) allows to exploit the enormous amount of computing power available in modern data centers, by decoupling the computing needs of the applications from physical processors and memory; moreover, VM are an efficient mean to neutrally save the state of a complex computation, to spawn different instances of the same computing environment or to implement safety and security related strategies and lower the overall risk in sensitive applications.

The drawback of using VM is their startup phase. When not in use, a VM is normally stored in the cloud storage system, with an (uncompressed) footprint that can be around several hundreds of megabytes. Additionally, a VM may use a persistent virtualized storage unit, that is logically mounted during the startup phase. Starting a VM, by using a standard predefined snapshot, or restarting a VM, previously stored as a snapshot at the end of the previous running period, is thus a time consuming task, due to data transfers from the storage subsystem of the cloud and the memory of the physical server chosen to run it.

As the schedule of the cloud depends on the workload, a snapshot (and, in case, its persistent storage) is not necessarily stored in the same node that can run it when needed (e.g.

OpenStack): a new VM instance from a standard image has to be retrieved from the image repository; an existing stored snapshot may need to be moved on the node that offers enough physical resources and time slots; if the architecture is made of different nodes for computing and storage, the snapshot obviously needs to be properly sent to the computing node; if the storage subsystem uses a distributed file system (e.g. CEPH), the snapshot retrieval involves even more complex mechanisms.

As a consequence, VM startup relies on the efficiency of the network layer, that is the part of cloud architectures that grows at the lowest pace. In this paper we present a modeling technique to evaluate the impact of the network layer and its organization on the VM startup time in high scale cloud architectures based on a three-tier network and a standard, replication based distributed storage model.

The paper is organized as follows: the next Section presents related works; Section III introduces the reference scenario for this work; Section IV describes the modeling approach; Section V shows the application to a case study; conclusions follow in the last Section.

II. RELATED WORKS

The use of VM is a classic technique, known since the mainframe era, to optimize the use of a large amount of resources by smartly sharing them between different, independent and isolated complete software stacks, by running different operating systems on virtualized hardware. A good, performance evaluation oriented introduction is provided by [1], that also offers a good historical perspective. VM offer a great flexibility in the management of cloud resources, that may give great benefits if a proper performance analysis driven tuning is implemented [2], as many performance influencing factors arise from the complexity of the architecture and must be taken into account [3] [4] [5] [6] [7].

The most critical performance factor in a modern data center is the efficiency of the network (at the point that data dependent structures may be needed [8]): to get an idea of the needs of a large data center, [9] reports about how this problem has been studied and what solutions have been implemented in

Google facilities. Several well spread architectures emerged, such as the three layer, the Clos, the fat tree [10] and the DCell ones [11] [12], but other solutions have been proposed as well (e.g. VL2 [13] and CONGA [14]). The problem is relevant also in distributed data centers [15] [16], but the scope of this paper is focused on the internal infrastructure of a single, high scale datacenter.

Performance evaluation studies on the main cloud network architectures have been performed by means of simulation: two very good examples are given by [11] and [12], that give a complete and comparative panorama. A simulative approach is potentially capable of allowing the analysis of very large scale clouds, at the cost of a long computational effort: simulation time is as much longer as much the system behaviors are variable and its scale is large. In this paper an analytical approach is preferred. An important issue is also the evaluation of energy consumption in cloud networks: this is out of the scope of this paper, but interested readers can find an interesting introduction and recent results in [17] and [18].

The authors already applied analytical and simulative methods to performance evaluation of cloud systems, both in small [19] [20] [21] [22] [23] and large scale [24] [25] [26] [27] [28]. Although analytical methods are known to be affected by the state space explosion problem, some approaches (e.g. product forms [29] and Markovian agents [30]) proved to be effective tools to overcome this limit. In this paper Markovian agents are exploited (as in [26], [27] and [28]) for their special suitability in modeling systems with very large number of states with increasing precision.

In this paper OpenStack cloud architecture is used as a reference. The management of VM images is documented at [31]; some technical information about typical VM images for OpenStack can be found at [32]. The network solicitation due to a VM is described at [33], while the integration of the operating system of a VM is described at [34] (using Ubuntu Linux distribution as an example). The main advantage of our approach with respect to the rest of the literature is the capability, thanks to the adoption of Markovian agents, for seamlessly scaling up the models (and the dimensions and complexity of datacenters) to thousands of components, while keeping an analytical approach and increasing the precision of the approximation.

III. SCENARIO

In this work we focus on a datacenter of medium or large scale. Figure 1 shows a simplified architecture of the considered scenario. In a datacenter, the IT equipment is enclosed into fixed form factor cases called *rack units*. Units include *computing servers*, *storage servers*, *network equipments* and *power supply units* (PSUs). In this work we will not focus on PSUs; network equipments include switches, routers, firewalls and load balancers: in this work we will only focus on switches. Units are organized in columns, that we will simply address as *racks*. Racks are further organized in corridors, to improve the air circulation and the cooling of units. In particular, corridors are organized into *cold aisles*

and *hot aisles*. The former ones present the front panels of the equipments, and allow technicians to access the controls of the units. The latter ones instead hold the backs of the units, and they are where cables interconnecting the units are placed. Cool air, produced by fans or air conditioners, enters the room from the cold aisles, flows through the units, cools them down, and exits the room from the hot aisles. Computing servers are usually special multiprocessor, multicore, and multithreaded power and network redundant x86 PCs. They are usually equipped with a relatively large amount of memory (currently in the range of 64-128 GB), and can run around 40-80 threads in parallel. They are however equipped with a limited disk space, and they relay to external storage to hold most of the persistent data. Storage units include both RAIDs (Rapid Array of Independent Disks) and JBOD (Just a Bunch Of Disks). The former are more expensive and require more advanced controllers: however they allow for both greater performance and reliability. The latter are much simpler and less expensive disk enclosures, whose task is just to allow computing units to mount them and use them as they were internal disks. Both units can be equipped with both rotational disks (HDD) or solid-state disks (SSD): usually a datacenter integrates all possible combinations of technologies to define different disk pools to be used for different purposes.

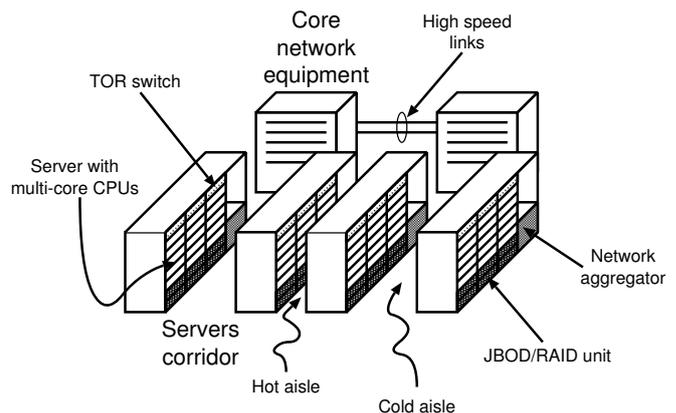


Fig. 1. Architecture of a datacenter.

Several network interconnection strategies for datacenters have been studied in the literature: a good survey can be found in [11]. In this work, we will mainly focus on the three-level network architecture as shown in Figure 2. Computing and storage units are directly connected to a switch that is defined as *Access switch* and that composes the so called *Access layer*. The switches can be positioned in two points that are usually addressed as *Top Of the Rack* (TOR) or *End Of the Line* (EOL). In the former, each rack has a switch, usually positioned in the top-most slot (for this reason it is called "top of the rack"). It has the advantage of requiring a small number of shorter cables. However it can reduce performances by placing additional bottlenecks, and it can reduce the size of the infrastructure. In the other topology, switches are put at the end of each line of racks. EOL allows a greater scalability, but it is

usually more expensive compared to the TOR solution. In our example we will focus on the TOR topology. Access switches are connected together using another level of switches, called *Aggregation layer*, that partitions the datacenter topology into a set of disjoint groups. In the example of Figure 1, aggregation switches are placed at the end of each corridor, and all the TOR access switches of the corresponding row of racks are connected to them. The connectivity of the datacenter is then completed by a further network, called the *Core layer*, that allows the communication between different aggregation switches. This organization however is affected by a problem known as the *bisection bandwidth*, which limits the maximum communication speed among nodes connected at different sides of the considered switch. Two techniques can be used to increase the available bisection bandwidth. Links that interconnect the different layers together might be characterized by different network technologies that might result in different link speeds. Since as the layer increases from access to aggregate, and from aggregate to core, the number of connected nodes increases as well, the bisection bandwidth can be increased by using faster communication technologies for links at higher levels. The second common way to increase the bisection bandwidth is by adding extra switches at each access layer, and using protocols like Equal Cost Multi-Path (ECMP) [35] to equally share the traffic among the different routes. For example, Figure 2 shows a 36 nodes architecture where each access switch is connected to three nodes (two for computing and one for storage), and access switches are grouped into bunches of three by the aggregation layer. Finally the four groups are connected together with the core layer. The bisection bandwidth is increased by using two aggregation layers per switch, and the by having three core layer switches.

In this work we mainly focus on a cloud datacenter, where all the computing nodes are used to host VMs. In our scenario, users are IaaS clients, that require the system to provide them a VM. Each user will then use the VM to run his software, and release it after use. As in a classical cloud scenario, VMs are started from *images*, that contain the filesystem of the OS plus all the other software that could be run in the VM. Persistent data are then stored using special block services set up by the cloud provider: they usually simulate the presence of a network connected disk that can be reached using the iSCSI protocol (a specific protocol that encapsulates SCSI data inside internet packets). For example, in *Openstack* [36], images are stored by a service called *Glance*, while persistent storage is provided by another service called *Cinder*. Both services use a lower-level block storage service (which in Openstack is called *Swift*). This file system architecture creates a high load over the network, which in many occasions becomes the real bottleneck of the system. In particular, the lifetime cycle of a VM, together with its storage access, is shown in Fig. 3. Initially, VM OS root disk images and persistent volume storage images are divided into blocks that are spread over the storage nodes of the datacenter (Fig. 3a). Root disk images size ranges from few tenth of MBs (for the smallest OS distributions) to several tenth of GB (for Windows based

OS, or for more featured Linux installations). As soon as a VM is started, its root disk image is copied into a local drive of the computing node where the VM is run (Fig. 3b). This creates a strong utilization of the network, since GBs of data must be transferred inside the datacenter. After the image has been copied, the virtual machine manager can start the VM. Each OS running on a VM usually can access at least three different disks: the root disk that contain the OS, a fast but small local disk (called the *ephemeral storage*), and a remote persistent storage. The main characteristic of ephemeral disks is that they are not persistent: when the VM is released, they are cleaned, and all their content is lost. During normal operations, the VM accesses the locally connected disks: the root disk to install OS updates or other software that must be run on the VM; ephemeral disks to hold temporary data. In this case the network is usually accessed only to access the persistent storage (Fig. 3c). Even if the exact access pattern is cloud-architecture dependent, it is usually performed by locally caching the data, and only relatively large blocks of packed data are sent across the network. At the release of the VM, the resources required to hold both the root and the ephemeral disks must be released. The user might require to perform a snapshot of the root disk in order not to lose the OS updates that have been made during the VM execution. This process is called *shelving* in Openstack terminology, and it requires that the new disk image must be transferred from the node that is releasing the VM to a storage node (Fig. 3d).

IV. MODELING APPROACH

Markovian Agents [37] are a formalism used to describe large system where elements can interact. Such models are solved using Mean Field Analysis [38]. In this case, agents do not communicate using messages, as ordinary Markovian Agents do, but they can influence each other via induction: the rate of jumping from one state to another can be influenced by the number of agents in a given state at a given location. Moreover, agents can increase in number or decrease (either spontaneously or induced by other agents), or they can multiply during the transitions.

The Markovian Agent based model depicted in Fig. 4 describes the behavior of a compute node. The system receives a total of Λ requests for activation of new VMs per time unit. Each node i receives requests at rate $\lambda_i(\Pi)$ (with $\sum_i \lambda_i(\Pi) = \Lambda$) where Π represents the count of agents in each state for each node. In particular, VMs are randomly assigned to nodes, with a probability that is proportional to the number of free VMs. Let us call $free_i(\Pi)$ the number of VMs that can still be assigned to node i , and let us call $\lambda_i(\Pi)$. We then have:

$$\lambda_i(\Pi) = \Lambda \frac{free_i(\Pi)}{\sum_j free_j(\Pi)}.$$

If the disk image of the starting VM is locally present, the agent goes in state *Local* with probability $p_{inCache}$ to simulate the immediate start of the computation. Otherwise, the agent goes in state *Startup_{ij}* to represent the image transfer from storage node j to compute node i . The image is transferred at

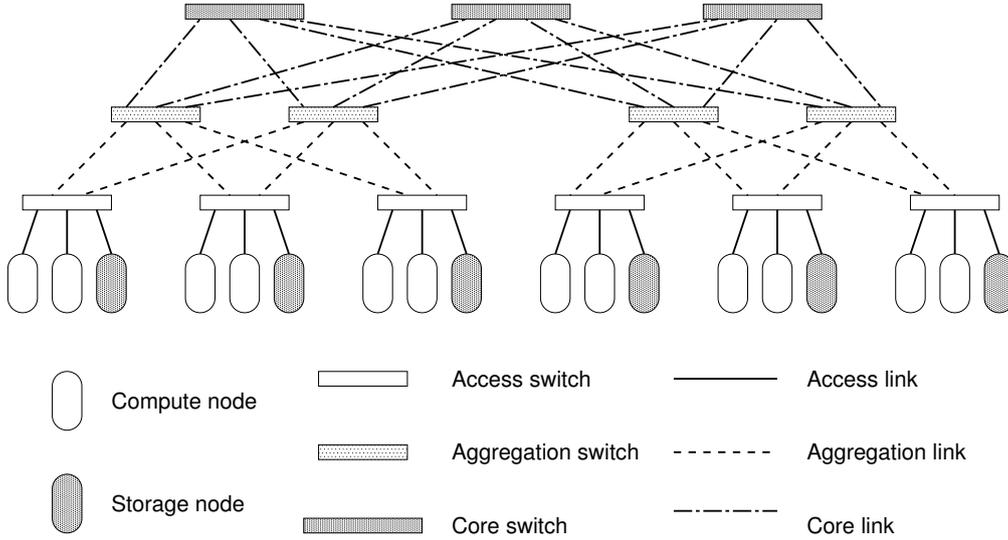


Fig. 2. Logical architecture of a three-tier datacenter interconnection network.

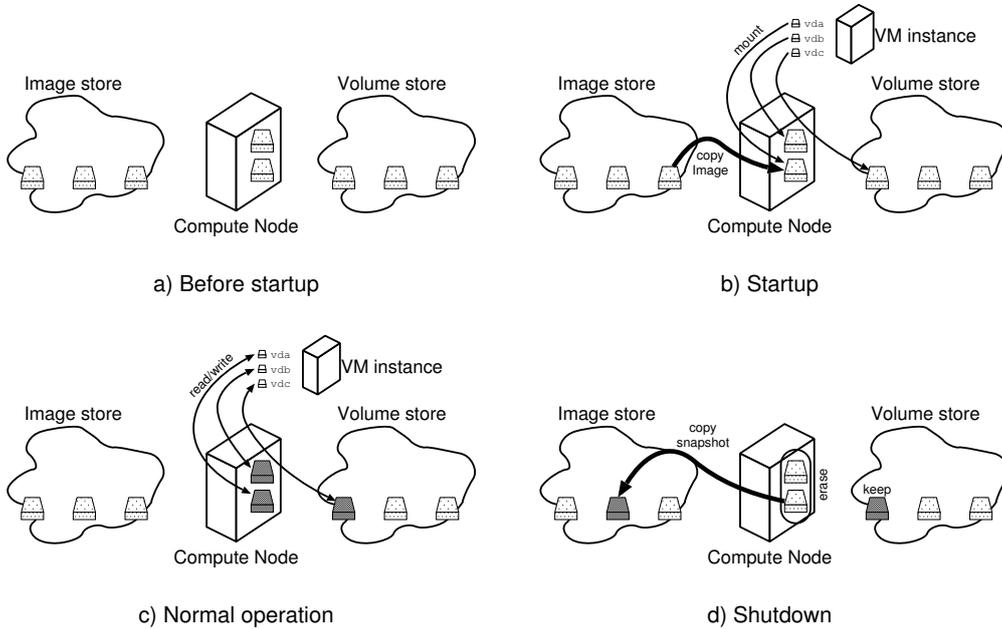


Fig. 3. Storage access during the lifetime of a VM: a) storage organization prior to startup, b) startup phase, c) disk access during normal operation, d) shutdown procedure.

rate $\sigma_{Startup_{ij}}$, that is equal to the speed of the link performing as bottleneck in the route connecting the compute node i to the storage one j . To be more precise, the computation of the speed of the route is computed in this way:

- 1) the total number of VMs $N_{R_{ij}}$ transferring data from each compute node i to each storage node j is computed;
- 2) let us call $\mathcal{R}_{l_k} = \{R_{ij}, \dots\}$ the set of all routes R_{ij} that traverses a link l_k . The total number of VMs N_{l_k} using link l_k is computed by considering all the possible routes ij that traverses that link, and multiplied by the sharing factor $sh(R_{ij}, l_k)$ of that link in the communication: $N_{l_k} = \sum_{R_{ij} \in \mathcal{R}_{l_k}} N_{R_{ij}} \cdot sh(R_{ij}, l_k)$. Sharing factor $sh(R_{ij}, l_k)$ allows

to model protocols like the ECMP;

- 3) for each link l_k , actual link speed σ_{l_k} is determined as $\sigma_{l_k} = \frac{C_{l_k}}{\max(N_{l_k}, 1)}$, where C_{l_k} is the maximum effective speed of link l_k measured in MB/s;
- 4) for each route R_{ij} , route speed $\sigma_{R_{ij}}$ is computed as the minimum capacity along the path: $\sigma_{R_{ij}} = \min_{l_k \in \mathcal{L}_{R_{ij}}} \sigma_{l_k}$, where $\mathcal{L}_{R_{ij}} = l_k$ is the set of link used by route R_{ij} ;
- 5) let us call $D_{VMimage}$ the average size of a VM image.

Rate $\sigma_{Startup_{ij}}$ is then determined as $\sigma_{Startup_{ij}} = \frac{\sigma_{R_{ij}}}{D_{VMimage}}$.

When the transfer is completed, the agent goes in state *Local*. The VMs session has duration $1/\mu_{shutdown}$; once it is

terminated the agent goes in state *Shutdown* with probability $1 - p_{noShelve}$ to model the user has performed the shelve action, otherwise the agent leaves the system. During the session, access to the remote disk can be requested by some applications at rate $\mu_{BlockIO}$. This behavior is modeled by the agent moving to the state *Block I/O_{ij}*. The agent returns to the normal state when the I/O is completed. This occurs at rate $\sigma_{BlockIO_{ij}} = \frac{\sigma_{R_{ij}}}{D_{Block}}$, where D_{Block} is the average I/O block size. If the VM requests to shelve the new image, the time spent to leave the system with rate $\sigma_{Shutdown}$ considers the copy of the image snapshot of size $D_{VMsnapshot}$. Also in this case the transfer speed depends on the bottleneck link between the computation and the storage nodes, and it is defined as $\sigma_{Shutdown} = \frac{\sigma_{R_{ij}}}{D_{VMsnapshot}}$

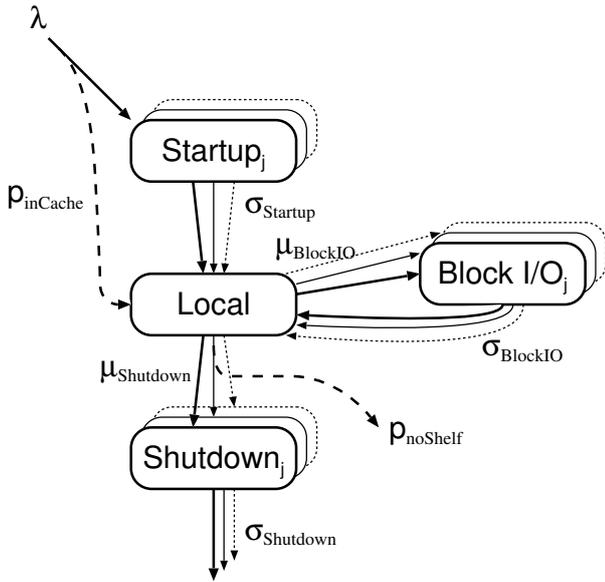


Fig. 4. Agent model of a compute node.

V. A CASE STUDY

To test the methodology, we apply the model proposed in Section IV to study the performances of four different patterns for placing computing and storage nodes inside a datacenter. In particular the considered scenarios are:

- **Distributed storage** (Fig. 5a). In this case there are no specific storage nodes, and disks are held directly inside the computing node. In this scenario, computing nodes have a slightly more limited capacity in term of VMs they can run, since they must also handle part of the storage requests.
- **Storage on rack** (Fig. 1). Each rack has its own set of storage nodes. This means that computing node might reach part of the storage nodes using just the switch at the corresponding access layer.
- **Storage on aisle** (Fig. 5b). In each aisle, there is a rack that includes all the storage nodes. In this way,

computing nodes must use switches at the aggregation layer to reach their storage units. However they can be easier to maintain, since disks are located in a limited number of locations (i.e. one per aisle).

- **Storage in a given area** (Fig. 5c). Storage nodes are concentrated in a specific aisle (which might also be physically located in a different room with respect to the computing nodes). This has the disadvantage that storage nodes can be accessed only passing through the core layer. However it ensures a higher security, allowing the storage to be located in different places.

All the scenarios share the same number of nodes $N_{nodes} = 36$, and the same maximum number of VMs that can be run on the infrastructure $N_{VMs} = 1920$. Scenarios 2, 3 and 4 are characterized by $N_{compute} = 24$ computing nodes, $N_{storage} = 12$ storage nodes, and each node has the capacity of running up to $N_{VMs} \times Node = 80$ VMs. In scenario 1, all nodes act both as computing and storage device: for this reason their capacity of running VMs $N_{VMs} \times Node$ has been reduced accordingly to keep the maximum capacity of the system $N_{VMs} = 1920$ as in the other scenarios. Links are characterized by the following speed: $C_{Access} = 500$ Mb/s at the access layer, $C_{Aggr} = 500$ Mb/s at the aggregation layer and $C_{Core} = 500$ Gb/s at the core layer¹. The average VM image size has been set to $D_{VMimage} = 80$ GB, while the snapshot size for VMs that are shelved (i.e. the difference from their original image) has been set to $D_{VMsnapshot} = 50$ MB. VMs perform block I/O on the average $\mu_{BlockIO} = 1$ block/h, and the block size is $D_{Block} = 10$ MB. Requests of new VMs activations arrive at a rate varying in the range $\Lambda = 5..30$ req./h, and each VM has an average duration of $1/\mu_{shutdown} = 25$ h. The probability of having a VM in cache is $p_{inCache} = 0.1\%$, and the probability of not shelving a terminating VM is $p_{noShelve} = 90\%$.

Figures 6, 7 and 8 show the average, minimum and maximum utilization of the links respectively at the access, aggregation and core layer. At the access layer, Scenarios 3 and 4 have a higher utilization since nodes produce a higher traffic to obtain the VMs due to the distribution of storage nodes. At the access layer, the only one having a lower utilization is Scenario 4, that is also the only one producing traffic at the core layer. In this case, however, the smaller utilization is due to the fact that the system saturates and creates a bottleneck for some nodes at the access layer.

Figures 9, 10 and 11 show the number of free VMs, the number of VMs performing IO operations (i.e. VMs laying in Startup, BlockIO, and Shutdown states), and the number of VMs in normal activities, respectively. As explained above, Scenarios 1 and 2 are more stable, as a consequence of the distribution of computing and storage nodes. It follows that they have a higher number of VMs performing the normal activity whereas in Scenarios 3 and 4 there is a higher number of VMs executing IO operations.

¹These speeds roughly corresponds to the maximum effective throughput that can be achieved on standard 1GB and 10GB Ethernet technologies

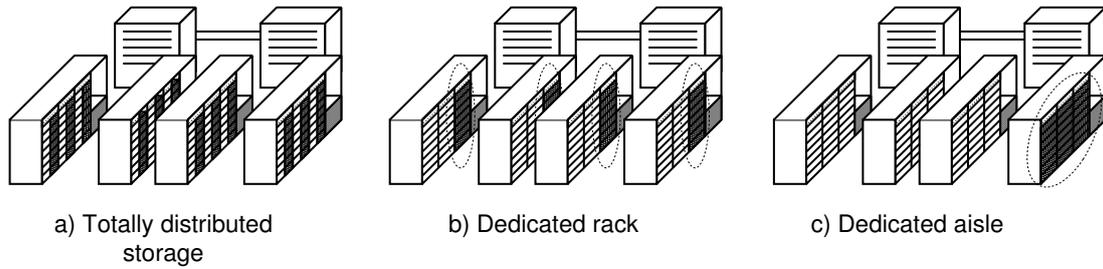


Fig. 5. Three alternative storage device organizations: a) storage is co-located with the computing nodes, b) storage is on a dedicated rack, c) storage is on a dedicated aisle.

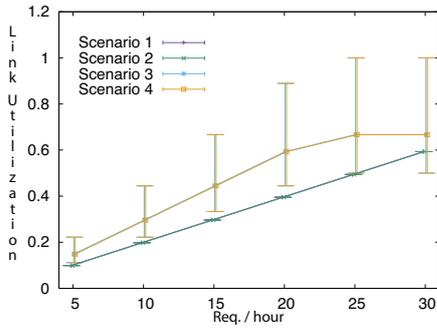


Fig. 6. Utilization of the links at access layer

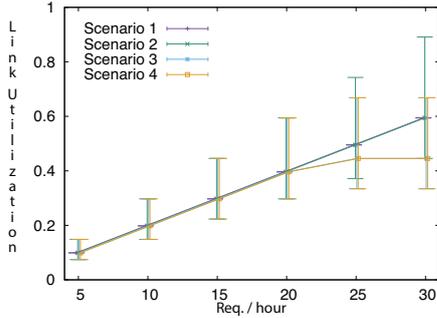


Fig. 7. Utilization of the links at aggregation layer

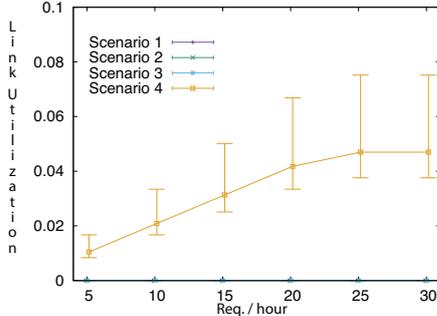


Fig. 8. Utilization of the links at core layer

VI. CONCLUSIONS

In this paper we proposed an approach for performance evaluation of the effects of networks in clouds. Our results, at the best of our knowledge, allow researchers and practitioners to model higher scale cloud systems with respect to previous literature, including architectures composed of more than one

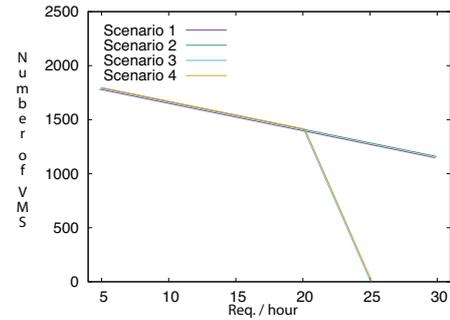


Fig. 9. Free VMs

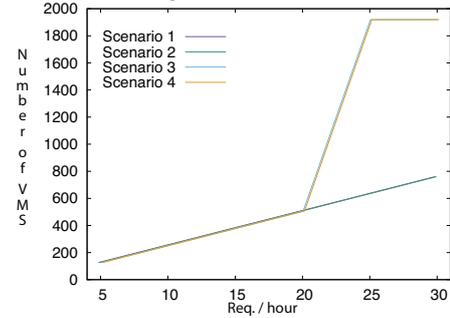


Fig. 10. VMs performing IO operations (Startup-Block-Shutdown)

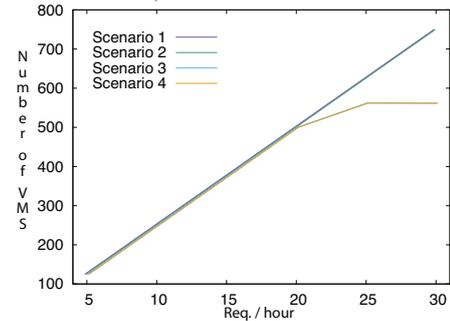


Fig. 11. VMs performing normal activity

data center. Future works include the integration with our previous proposals for a detailed overall evaluation of all aspects of cloud systems. Moreover, we will study other network topologies that rely on commodity hardware such Fat-tree organizations, to see if alternative to the three layer

architectures could improve the performance and reduce the cost of a data-center.

REFERENCES

- [1] D. A. Menascé, "Virtualization: Concepts, applications, and performance modeling," in *Proc. of The Computer Measurement Groups 2005 International Conference*, 2005.
- [2] 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.
- [3] N. Huber, M. Von Quast, F. Brosig, and S. Kounev, "Analysis of the performance-influencing factors of virtualization platforms," in *Proc. 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.
- [4] B. J. Watson, M. Marwah, D. Gmach, Y. Chen, M. Arlitt, and Z. Wang, "Probabilistic performance modeling of virtualized resource allocation," in *Proc. of the 7th international conference on Autonomic computing*, ser. ICAC '10. NY, USA: ACM, 2010, pp. 99–108.
- [5] 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.
- [6] M.-A. Vasile, F. Pop, R.-I. Tutueanu, V. Cristea, and J. Koodziej, "Resource-aware hybrid scheduling algorithm in heterogeneous distributed computing," *Future Generation Computer Systems*, vol. 51, pp. 61 – 71, 2015.
- [7] A. Sfrént and F. Pop, "Asymptotic scheduling for many task computing in big data platforms," *Information Sciences*, vol. 319, pp. 71 – 91, 2015.
- [8] U. Fiore, F. Palmieri, A. Castiglione, and A. De Santis, "A cluster-based data-centric model for network-aware task scheduling in distributed systems," *International Journal of Parallel Programming*, vol. 42, no. 5, pp. 755–775, 2014.
- [9] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, "Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 183–197, Aug. 2015.
- [10] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, Aug. 2008.
- [11] K. Bilal, S. U. Khan, L. Zhang, H. Li, K. Hayat, S. A. Madani, N. Min-Allah, L. Wang, D. Chen, M. I. Iqbal, C. Xu, and A. Y. Zomaya, "Quantitative comparisons of the state-of-the-art data center architectures," *Concurrency and Computation: Practice and Experience*, vol. 25, no. 12, pp. 1771–1783, 2013.
- [12] R. D. S. Couto, S. Secci, M. E. M. Campista, and L. H. M. K. Costa, "Reliability and survivability analysis of data center network topologies," *CoRR*, vol. abs/1510.02735, 2015.
- [13] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "V12: A scalable and flexible data center network," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 51–62, Aug. 2009.
- [14] M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut, V. T. Lam, F. Matus, R. Pan, N. Yadav, and G. Varghese, "Conga: Distributed congestion-aware load balancing for datacenters," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 503–514, Aug. 2014.
- [15] F. Palmieri, U. Fiore, S. Ricciardi, and A. Castiglione, "Grasp-based resource re-optimization for effective big data access in federated clouds," *Future Generation Computer Systems*, vol. 54, pp. 168–179, 2016.
- [16] S. Spoto, M. Gribaudo, and D. Manini, "Performance evaluation of peering-agreements among autonomous systems subject to peer-to-peer traffic," *Perform. Eval.*, vol. 77, pp. 1–20, 2014.
- [17] C. Fiandrino, D. Kliazovich, P. Bouvry, and A. Zomaya, "Performance and energy efficiency metrics for communication systems of cloud computing data centers," *IEEE Trans. on Cloud Computing*, vol. PP, no. 99, pp. 1–1, 2015.
- [18] P. Ruiu, A. Bianco, C. Fiandrino, P. Giaccone, and D. Kliazovich, "Power comparison of cloud data center architectures," in *Proc. of the 2016 IEEE International Conference on Communications (ICC)*, 2016.
- [19] E. Barbierato, M. Gribaudo, and M. Iacono, "A performance modeling language for big data architectures," in *ECMS, W. Rekdalsbakken, R. T. Bye, and H. Zhang, Eds. European Council for Modeling and Simulation*, 2013, pp. 511–517.
- [20] —, "Performance evaluation of NoSQL big-data applications using multi-formalism models," *Future Generation Computer Systems*, vol. 37, no. 0, pp. 345–353, 2014.
- [21] —, "Modeling apache hive based applications in big data architectures," in *Proc. of the 7th International Conference on Performance Evaluation Methodologies and Tools*, ser. ValueTools '13. ICST, Brussels, Belgium: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2013, pp. 30–38.
- [22] D. Cerotti, M. Gribaudo, M. Iacono, and P. Piazzolla, "Modeling and analysis of performances for concurrent multithread applications on multicore and graphics processing unit systems," *Concurrency and Computation: Practice and Experience*, pp. n/a–n/a, 2015.
- [23] —, "Workload characterization of multithreaded applications on multicore architectures," in *ECMS, Proc. of the 28th European Conference on Modelling and Simulation, ECMS 2014, Brescia, Italy, May 27-30, 2014*. European Council for Modeling and Simulation, 2014, pp. 480–486.
- [24] M. Gribaudo, M. Iacono, and D. Manini, "Improving reliability and performances in large scale distributed applications with erasure codes and replication," *Future Generation Computer Systems*, vol. 56, pp. 773 – 782, 2016.
- [25] —, "Modeling replication and erasure coding in large scale distributed storage systems based on CEPH," in *Proc. of XII conference of the Italian chapter of AIS*, ser. Lecture Notes in Information Systems and Organisation. Springer Berlin / Heidelberg, 2016, vol. to appear.
- [26] A. Castiglione, M. Gribaudo, M. Iacono, and F. Palmieri, "Exploiting mean field analysis to model performances of big data architectures," *Future Generation Computer Systems*, vol. 37, no. 0, pp. 203–211, 2014.
- [27] —, "Modeling performances of concurrent big data applications," *Software: Practice and Experience*, vol. 45, no. 8, pp. 1127–1144, 2015.
- [28] E. Barbierato, M. Gribaudo, and M. Iacono, "Modeling and evaluating the effects of Big Data storage resource allocation in global scale cloud architectures," *International Journal of Data Warehousing and Mining*, vol. 12, no. 2, pp. 1–20, 2016.
- [29] E. Barbierato, G.-L. D. Rossi, M. Gribaudo, M. Iacono, and A. Marin, "Exploiting product forms solution techniques in multiformalism modeling," *Electronic Notes in Theoretical Computer Science*, vol. 296, no. 0, pp. 61 – 77, 2013.
- [30] M. Gribaudo and M. Iacono, "A different perspective of agent-based techniques: Markovian agents," in *Intelligent Agents in Data-intensive Computing*, ser. Studies in Big Data, J. Kolodziej, L. Correia, and J. Manuel Molina, Eds. Springer International Publishing, 2016, vol. 14, pp. 51–70.
- [31] "OpenStack: Images and instances," <http://docs.openstack.org/admin-guide-cloud/compute-images-instances.html>, accessed: 2016-02-06.
- [32] "Where to find OpenStack cloud images," <https://thornelabs.net/2014/06/01/where-to-find-openstack-cloud-images.html>, accessed: 2016-02-06.
- [33] "OpenStack networking tutorial: Single-host Flat-DHCPManager," <https://www.mirantis.com/blog/openstack-networking-single-host-flatdhcpmanager/>, accessed: 2016-02-06.
- [34] "Ubuntu documentation: CloudInit," <https://help.ubuntu.com/community/CloudInit>, accessed: 2016-02-06.
- [35] C. Hopps, "Analysis of an equal-cost multi-path algorithm," United States, 2000.
- [36] "OpenStack web site," <https://www.openstack.org/>, accessed: 2016-02-06.
- [37] M. Gribaudo, D. Cerotti, and A. Bobbio, "Analysis of on-off policies in sensor networks using interacting markovian agents," in *Proc. of the 4th International Workshop on Sensor Networks and Systems for Pervasive Computing - PerSens 2008*, 2008.
- [38] B. M. and L. J.Y., "A class of mean field interaction models for computer and communication systems," *Performance Evaluation*, vol. 65, no. 11–12, pp. 823–838, 2008.

A Multi-Formalism Framework to Generate Diagnostic Decision Support Systems

Giuseppe Cicala, Marco De Luca, Marco Oreggia, Armando Tacchella

KEYWORDS

Modeling and simulation of Cyber-Physical Systems, Knowledge-based modeling formalisms, Actor-based simulation.

ABSTRACT

The task of a Diagnostic Decision Support System (DDSS) is to deduce the health status of a physical system. In this paper, a multi-formalism framework to generate DDSS software based on formal descriptions of the application domain and the diagnostic computations is proposed. The key idea is to describe systems and related data with a *domain ontology*, and to describe diagnostic computations with an *actor-based model*. Implementation-specific code is automatically generated from such dual-formalism descriptions, while the structure of the DDSS is invariant across applications. An evaluation involving an artificial scalable domain related to the diagnosis of air conditioning systems is presented to exemplify and to test the proposed framework.

I. INTRODUCTION

Diagnostic Decision Support Systems (DDSSs) help humans in the deduction of information about the health status of some observed physical system. From a practical point of view, the availability of digital sensors, reliable and high-capacity networks and powerful processing units, makes automated diagnosis applicable to an increasing number of systems. However, data and diagnostic rules remain domain-dependent, and the implementation of a DDSS requires the development of substantial portions of ad-hoc software which can hardly be recycled. Indeed, while most of the existing literature about DDSS focuses on improving performances in some domain of interest, to the best of our knowledge there is no contribution in the way of generating customized DDSS from high-level specifications.

The research presented in this paper attempts to fill this gap by developing a framework to generate customized DDSSs using a multi-formalism approach. Multi-formalism modeling — see, e.g., [GI13] — refers to tools and techniques wherein several different formalisms are exploited to achieve a specific goal. The

G. Cicala, M. Oreggia and A. Tacchella are with “Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi” (DIBRIS), University of Genoa, Viale Causa 13, 16145 Genoa, Italy, E-mail: name.surname@unige.it — M. De Luca is with ABIRK Italia S.r.l., Corso MonteGrappa 1/1A, 16137 Genoa, Italy, E-mail: marco.deluca@abirk.com

combination of formalisms is useful whenever specifying a system with a single modeling language would be hard, if not impossible. In this paper, two “classical” AI formalisms are combined to generate DDSSs: systems are described with ontologies in the sense of [Gru95], i.e., “formal and explicit specification of conceptualizations”; diagnostic computations are described with actor-based models as introduced in [Agh85] with the extensions found in [LTSS11].

More in detail, the choice of ontologies is motivated by their increasing popularity outside the AI community — mainly due to Semantic Web applications — and the added flexibility that they provide over traditional relational data models, e.g., the ability to cope with taxonomies and part-whole relationships, and the ability to handle heterogeneous attributes. It should be noticed that other proposals exist in the literature to extend the basic relational data model in order to handle more expressive domains, e.g., [CM94]. We consider ontologies because they provide a general-purpose, logically well-founded extension which also enjoys widespread use. Since it is expected that large quantities of data should be handled to provide meaningful input to the DDSS, the choice of the ontology language should be restricted to those designed for tractable reasoning like the DL-Lite family introduced by [CGL⁺05]. The choice of actor-based models is motivated by support for heterogeneous modeling, i.e., a situation wherein different parts of a system have inherently distinct properties, and therefore require different types of models. DDSSs are no exception to this pattern, since they are required to monitor and diagnose the behavior of heterogeneous systems, and they are themselves a composition of physical processes and computational elements.

Following the approach outlined above, a DDSS generator — called DiSeGnO for “Diagnostic Server Generation through Ontology” — has been developed. DiSeGnO outputs a DDSS given a formal description of the application domain — the *domain ontology* — and associated diagnostic computations — the *diagnostic rules*. DiSeGnO interprets such dual-formalism descriptions by generating a relational database from the domain ontology and then computing diagnostic rules using PTOLEMY II [EJL⁺03], an open-source software supporting experimentation with actor-based design. The generated DDSS is also wrapped by automatically generated web services which connect to the plant on one side, and to diagnostic dashboards on the other. The

conversion of the ontology design to a database structure is a key element in DiSeGnO. It preserves the high level description but, at the same time, it ensures quick access to data and leverages industry-standard database systems. The usage of PTOLEMY II as a rule engine enabled fast-prototyping of DiSeGnO, and might be replaced by a diagnostic rule compiler in practical applications. However, as the experimental analysis herein presented shows, even in its current PTOLEMY II-based implementation, DiSeGnO can process a substantial flow of data from an incoming (simulated) plant in real-time. In this sense, our work is similar in spirit to [FMMV16], as both contributions propose to merge different formalisms in order to describe complex systems properly.

The rest of the paper is structured as follows. In Section II an introduction to ontologies and actor-based models is given. A case study about Heating, Ventilation and Air Conditioning (HVAC) systems in households is presented in Section III. In Section IV the architecture of DiSeGnO and the main components to generate DDSSs are presented. Finally, Section V shows the experimental evaluation of DiSeGnO on the HVAC case study. The paper is concluded in Section VI with some final remarks and an outline of a future research agenda.

II. BACKGROUND

Ontology-based data access (OBDA) relies on the concept of *knowledge base*, i.e., a pair $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ where \mathcal{T} is the *terminological box* (*Tbox* for short) specifying the intensional knowledge, i.e., known classes of data and relations among them, and \mathcal{A} is the *assertional box* (*Abox* for short) specifying the extensional knowledge, i.e., factual data and their classification. Filling the Abox with known facts structured according to the Tbox is a process known as *ontology population*. One of the mainstream languages for defining knowledge bases is OWL 2 (Web Ontology Language Ver. 2) described in [MPSP⁺09]. Since OWL 2 is a World Wide Web Consortium’s recommendation, it is supported by several ontology-related tools. However, the logical underpinning of OWL 2 is the description logic *SRQIQ* whose decision problem is 2NEXPTIME-complete according to [Kaz08]. This makes the use of the full expressive power of OWL 2 prohibitive for an application like the one we are considering.

To retain most of the practical advantages of OWL 2, but to improve on its applicability, Motik et al. introduced *OWL 2 profiles* – see [MPSP⁺09]. Formally, an OWL 2 profile is a sub-language of OWL 2 featuring limitations on the available language constructs and their usage. In particular, the OWL 2 QL profile is described in the official W3C’s recommendation as “[the sub-language of OWL 2] aimed at applications that use very large volumes of instance data, and where query answering is the most important reasoning task.”. Given our application domain, OWL 2 QL is more appealing than both OWL 2 and other profiles, because it guarantees that conjunctive query answering and the consistency of the ontology can be evaluated efficiently.

OWL 2 QL logic underpinning is given by *DL-Lite_R*, one of the members of the *DL-Lite* family [CGL⁺05]. A detailed description of *DL-Lite_R* can be found in [CGL⁺05]. The most important feature of OWL 2 QL in this context is that, using the mapping techniques introduced in [RMC12], it is possible to keep the terminological view to reason about data, while storing the Abox elements as records in a relational database. Formally, given a knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, it is possible to build a database with a set of relations (tables) $R_{\mathcal{K}}$ such that the query $\mathcal{K} \models \alpha$ can be translated to a relational algebra expression over $R_{\mathcal{K}}$ returning the same result set. The choice of OWL 2 QL guarantees that the mapping from \mathcal{K} to $R_{\mathcal{K}}$ is feasible, and the translation of ontology-based queries to SQL queries will yield polynomially bounded expressions. In this way, it is possible to take the best of the two approaches, i.e., use ontologies to define the conceptual view of the domain, and databases to store actual data and connect to the other performances-critical elements of the generated DDSS, like data I/O and processing components.

The following notations and definitions are from [LTSS11]. Let S be a set of variables that take values in some universe \mathcal{U} . A *valuation* over S is a function $x : S \rightarrow \mathcal{U}$ that assigns to each variable $v \in S$ some value $x(v) \in \mathcal{U}$. The set of all assignments over S is denoted by \hat{S} . If $x \in \hat{S}$, $v \in S$ and $\alpha \in \mathcal{U}$, then $\{x \mid v \mapsto \alpha\}$ denotes the new valuation x' obtained from x by setting v to α and leaving other variables unchanged. *Timers* are a special type of variables that take values in \mathbb{R}_+ , i.e., non-negative real numbers. Let \mathbb{R}_+^∞ denote the set $\mathbb{R}_+ \cup \{\infty\}$, where ∞ denotes positive infinity. Finally, let $\perp \in \mathcal{U}$ and $\text{absent} \in \mathcal{U}$ denote “unknown” value or “absence” of a signal at a particular point in time, respectively.

An *actor* is a tuple $A = (I, O, S, s_o, F, P, D, T)$ where I is a set of *input variables*, O is a set of *output variables*, S is a set of *state variables*, and $s_o \in \hat{S}$ is a valuation over S representing the *initial state*; F is the *fire function*, defined as $F : \hat{S} \times \hat{I} \rightarrow \hat{O}$, that produces output based on input and current state; P is the *postfire function* defined as $P : \hat{S} \times \hat{I} \rightarrow \hat{S}$ that updates the state based on the same information of the fire function; D is a *deadline function* defined as $D : \hat{S} \times \hat{I} \rightarrow \mathbb{R}_+^\infty$ and T is a *time-update function* defined as $D : \hat{S} \times \hat{I} \times \mathbb{R}_+ \rightarrow \hat{S}$. It is assumed that F, P, D, T , are total functions, and I, O , and S are pair-wise disjoint. In the following, the terms *input*, *output* and *state* refer to valuations over I, O and S , respectively.

Every actor A defines a set of behaviors whose model is inspired by the semantic models of timed or hybrid automata. A *timed behavior* of A is a sequence

$$s_0 \xrightarrow{x_0/y_0} s'_0 \xrightarrow{x'_0/d_0} s_1 \xrightarrow{x_1/y_1} s'_1 \xrightarrow{x'_1/d_1} s_2 \xrightarrow{x_2/y_2} s'_2 \dots$$

where for all $i \in \mathbb{N}$, $s_i, s'_i \in \hat{S}$, $d_i \in \mathbb{R}_+$, $x_i \in \hat{I}$, $y_i \in \hat{O}$

$$y_i = F(s_i, x_i) \quad s'_i = P(s_i, x_i) \\ d_i \leq D(s'_i, x'_i) \quad s_{i+1} = T(s'_i, x'_i, d_i).$$

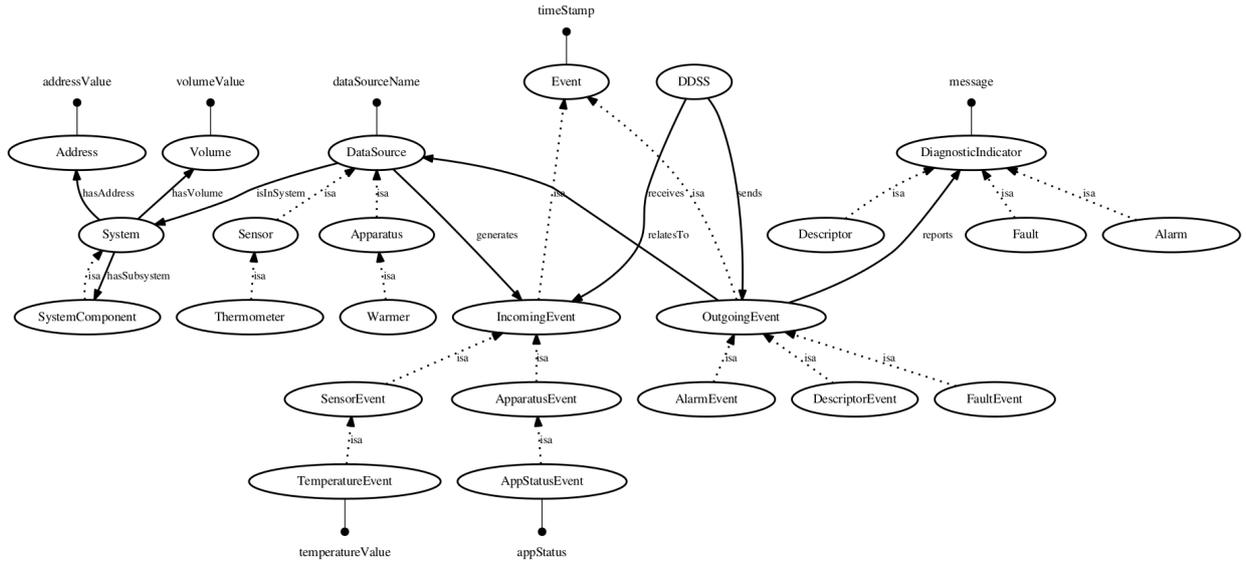


Fig. 1. Domain ontology for HVAC monitoring. Concepts are represented by ovals, concept inclusions (*is-a* relationships) are denoted by dashed arrows, roles are denoted by solid arrows, and attributes are denoted by dots attached to classes.

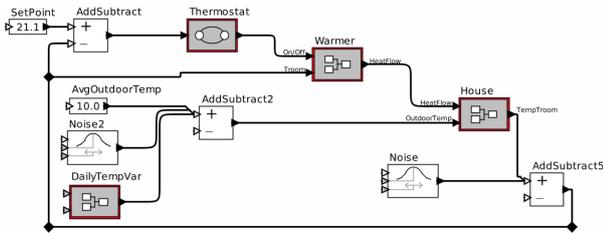


Fig. 2. Thermal model of a house with an HVAC unit sketched with PTOLEMY II graphical syntax. Gray boxes which include rectangles inside, e.g., House, are composite actors, whereas those with circles inside, e.g., Thermostat, are finite-state machines.

Intuitively, if A is in state s_i at some time $t \in \mathbb{R}_+$ and the environment provides input x_i to A , then A instantaneously produces output y_i using the fire function F , and moves to state s'_i using the postfire function P . The environment then proposes to advance time, but it does so “respecting” any restriction on the amount of time that may elapse. A “declares” such restrictions by returning a deadline $D(s'_i, x'_i)$. Next, the environment chooses to advance time by some concrete delay $d_i \in \mathbb{R}_+$, making sure that d_i does not violate the deadline provided by A . Finally, the environment notifies A that it advanced time by d_i , and A updates its state to s_{i+1} accordingly, using the time-delay function T .

III. HVAC CASE STUDY

HVAC systems are a classic topic in diagnostics — see, e.g., [NAL⁺07]. Here, the model¹ shown in Figure 2 is considered as an example. This model takes into account topology, thermal properties of materials, and warmer characteristics, i.e., temperature of output hot air and flow-rate. As shown in Figure 2,

¹The house thermal model can be downloaded from <http://www.mathworks.it/help/simulink/examples/thermal-model-of-a-house.html>

the main model components are Thermostat, Warmer and House subsystems. Thermostat allows fluctuations within a certain range above or below the desired set point. If House temperature drops below the set point minus allowed fluctuation, Thermostat turns on Warmer to provide a hot air flow at a constant rate and temperature. The heat flow is expressed by $\frac{dQ_{warmer}}{dt} = (T_{warmer} - T_{room}) \cdot M_{dot} \cdot c$ where $\frac{dQ}{dt}$ is the heat flow from Warmer to House, c is the heat capacity of air at constant pressure, M_{dot} is the air mass flow rate, T_{warmer} is the temperature of hot air and T_{room} is air temperature in the house. House subsystem calculates internal temperature variations. It takes into account the heat flow from Warmer and heat losses. Heat losses and the temperature time derivative are governed by $\frac{dQ_{losses}}{dt} = \frac{T_{room} - T_{outside}}{R_{eq}}$ and $\frac{dT_{room}}{dt} = \frac{1}{M_{air} \cdot c} \cdot \frac{dQ_{warmer}}{dt} - \frac{dQ_{losses}}{dt}$ where M_{air} is the mass of air inside House and R_{eq} is the equivalent thermal resistance of House. The DailyTempVar subsystem generates a daily fluctuations of outdoor temperature. Both inside and outside temperatures are affected by a Gaussian noise to simulate reading from real sensors.

The ontology of the HVAC domain is shown in Figure 1. The main concepts in the static part of the domain are System and DataSource. They are related by `isInSystem`, stating that every system has — possibly several — data sources attached to it. `hasSubsystem` relationship indicates that one System could be composed by one or more SystemComponent which are themselves subclasses of System. DataSource is the comprehensive class of elements that can generate diagnostic-relevant information. The main concepts in the dynamic part of the ontology are DDSS which receives instances of IncomingEvent and sends instances of OutgoingEvent. Notice that IncomingEvent

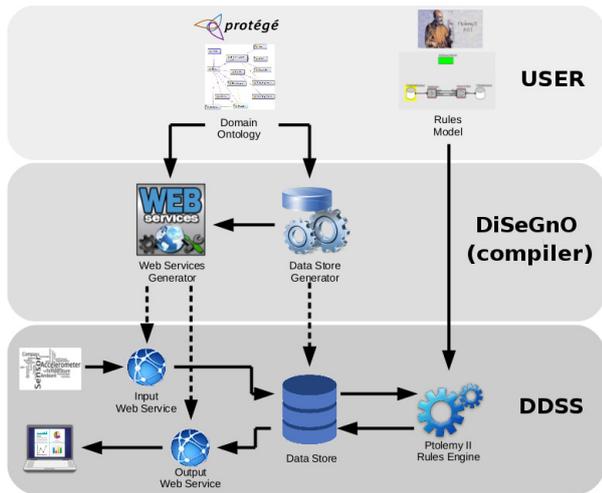


Fig. 3. Functional architecture and work-flow of the current DiSeGnO framework.

instances are connected to `DataSource` instances by the role `generates`, denoting that all incoming events, i.e., data from the observed system, are generated by some data source, i.e., some field sensor. Also every `OutgoingEvent` instance, i.e., every diagnostic event, `relatesTo` some instance of `DataSource`. This is because the end user must be able to reconstruct which data source(s) provided information that caused diagnostic rules to fire a given diagnostic event. `OutgoingEvent` specializes to `AlarmEvent`, `FaultEvent` and `DescriptorEvent`. Every `OutgoingEvent` instance is connected to one of `DiagnosticIndicator` instances, i.e. `Alarm`, `Fault` and `Descriptor` sub-concepts, by `reports` relation, in order to have a reference message about the diagnostic rules.

Diagnostic rules of interest have been extracted from the literature on HVAC systems — see, e.g., [RWLF04]. In particular, assuming that there is a single fault in the system at any time, air filter obstruction, thermostat fault and pressure loss in the compressor are investigated. In case of air filter obstruction, a reduction of air flow in output from the warmer results in a slow temperature drift away from the comfort zone. If the thermostat ceases to work properly, e.g., because its state is stuck to either on or off, then the house temperature stays permanently away from the comfort zone. In case of pressure loss in the compressor, a loss of refrigerant charge happens which diminishes the capability of the compressor. The domain ontology, as well as the model rules herein described are available on-line from

<http://www.aimslab.org/diseigno>

IV. DISEGNO FRAMEWORK

A. Software architecture

Figure 3 shows the current functional architecture and work-flow of DiSeGnO, organized in three phases. In the USER phase, the domain ontology and the rules model

are designed by the user. In the DiSeGnO phase, the system reads and analyzes both the domain ontology and the rules model. The output is code structured as shown in the DDSS phase. Here, input web services receive data from the observed physical system and record them in the generated data store. The rule engine feeds the diagnostic rules with records loaded from the data store and logs results, if any. Output web services can then be invoked to query the data store.

In the USER phase, the user is required to provide an ontology of the observed physical system which must be written using OWL 2 QL language. While this can be accomplished in several ways, the tool PROTÉGÉ [GMF⁺03] is suggested because it is robust, easy to use, and it provides, either directly or through plug-ins, several add-ons that facilitate ontology design and testing. The diagnostic computation model must be a sound actor diagram generated by PTOLEMY II which describes the processing to be applied to incoming data in order to generate diagnostic events. The only additional requirement on the rules model is that the set of external inputs of the diagram must coincide with the incoming events described in the ontology. Analogously, the set of external outputs of the diagram must coincide with the outgoing events.

The DiSeGnO phase contains the actual DDSS generation system which consists of two modules in the current implementation, namely the `Data Store Generator` and the `Web Services Generator`. Given the domain ontology, a data store is generated to record data and events. The data store is a relational database which is obtained by mapping the domain ontology to suitable tables. The web services generator creates services whose interface asks for incoming events of the correct type (input web services) and services which can be queried to obtain diagnostic events (output web services). Currently, the working prototype uses PTOLEMY II internal engine to run the rules model as if they were code run on top of an interpreter. This solution is straightforward to implement, but has the disadvantage of being potentially slow for real-world applications.

In the DDSS phase, the customized DDSS runs in a loop wherein (i) data is acquired from the observed system and stored in the internal database, (ii) the rules engine processes data and generates diagnostic events which are recorded on the database, and (iii) diagnostic data is served to end-user application. The details of the data acquisition on the observed system are not of concern to the DDSS generated by DiSeGnO, because it is the responsibility of the observed system control logic to implement the data acquisition part. This choice effectively isolates the physical details of data acquisition from the rest of the DDSS. Similarly, the generated DDSS is not concerned with the details of displaying and representing diagnostic data, because these data are made available through output web-services and it is responsibility of the user applications to read such data and present them in a meaningful way.

```

procedure VISITONTOLOGY(onto, d, r)
  g ← new Graph()
  VISITONTOLOGYREC(onto, onto.getThing(), g, d, r)
  return g
end procedure

procedure VISITONTOLOGYREC(onto, c, g, d, r)
  for all Concept s ∈ c.getSubConcepts() do
    father ← NIL
    if c ≠ onto.getThing() then
      father ← g.getTable(c)
    end if
    T ← g.getTable(s)
    if T is NIL then
      T ← new Table(s)
      for all Attribute a ∈ d.getDataAttribute(s) do
        T.addAttribute(a)
      end for
      g.addNode(new Node(T))
      if father is not NIL then
        r.add(new Relationship(T, father, '1 to n'))
        g.addEdge(T, father)
      end if
      VISITONTOLOGY(onto, s, g, d, r)
    else
      if father is not NIL then
        r.add(new Relationship(T, father, '1 to n'))
        g.addEdge(T, father)
      end if
    end if
  end for
end procedure

```

Fig. 4. Main algorithms of the Data Store Generator component.

B. From ontology to Database

Ontology has to be divided into two interconnected parts, namely a *static* and a *dynamic* part. In the static part, the ontology should contain a description of the observed physical system including entities for each relevant (sub)system and relationships among them. This part, once populated with the actual systems to be observed, does not require further updates while monitoring. On the other hand, the dynamic part describes *events*, including both the ones generated by the observed system and its components, and those output by the DDSS. An event is always associated to a timestamp, i.e., the time at which the event happens. Data associated to events can be of heterogeneous types, but we always distinguish between two kinds of events, i.e., those *incoming* to the DDSS from the observed system, and those *outgoing* from the DDSS. This distinction is fundamental, because DiSeGnO must know which events have to be associated with input and output web services, respectively. Furthermore, both events should be associated with the data sources, i.e., the elements of the static part which generate events or influence the generation of a diagnostic event.

As mentioned in the Introduction, the creation of a relational database from the ontology, i.e., the Tbox \mathcal{T} , allows efficient storage of the corresponding Abox \mathcal{A} . The knowledge base $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ can still be queried seamlessly, e.g., by using the mapping techniques described in [RMC12]. The algorithm used by DiSeGnO to encode an OWL 2 QL ontology into the structure of a relational database reads the ontology model from an OWL file into the internal representation *onto*; then it parses *onto* and extracts the map *dataMap* between

concepts and datatype properties and the map *relMap* between concepts and object properties (roles). At this point, it can visit the ontology by traversing the concept hierarchy with the function VISITONTOLOGY — see Figure 4 — and it creates the graph *dbGraph* containing part of the relational model corresponding to *onto*, using *dataMap* as *d* and *relMap* as *r*. Finally, it builds the relational model by considering all the relationships, and translates it into a database, considering all the nodes of *dbGraph* and building corresponding tables and constraints.

In more detail, VISITONTOLOGY and its sister procedure VISITONTOLOGYREC — see Figure 4 — perform a visit of the concept hierarchy contained in *onto* to create a corresponding graph stored in *dbGraph*. Since the concept hierarchy forms, by definition, a directed acyclic graph, a simplified implementation of depth-first search visit is sufficient to explore *onto* exhaustively. Inside VISITONTOLOGYREC a new table *T* — and a corresponding node in the graph *g* — is created for each concept contained in *onto*. Furthermore, all the datatype properties corresponding to the concept of *T* are retrieved from the map *d* and added to *T*. These will become attributes of the entity corresponding to the concept in the final relational database. Notice that a one-to-many relationship corresponding to the inheritance relation is added to *r*, the set of relationships extracted considering object properties in *onto*. As long as *d* is implemented with a constant-time access structure, the running time of VISITONTOLOGY is linear in the size of *onto*.

C. Rules Engine in Ptolemy

Database connection is guaranteed by a *DatabaseManager* actor that opens a connection and passes it to all actors accessing the database. Data are collected using generic *DatabaseQuery* actors that query the database via the specified *DatabaseManager* and provide results as arrays of records. Collected data are provided to other actors in the rule models according to their time stamp. Fault-detection rules are implemented in PTOLEMY II models using data-driven techniques. In particular, both rule 1 and 3 leverage Artificial Neural Networks (ANN) trained with Encog [Hea14] software. In the rule detecting fault 1, the ANN is used to estimate the value (in percentage) of air which is getting through the warmer. In case of fault 3, the ANN estimates the value of gas pressure in the compressor circuit. Both estimations are required because no physical sensors are available to measure those quantities directly — ANNs act as *virtual sensors* as in [KPJ06]. Rule 2 is based on a statistical outliers detection on the population of time intervals between thermostat switching cycles. Outliers are identified by a finite state machine that assesses whether or not they fall within a set of numerical boundaries called *fences*. If the time interval between two consecutive warmer “on” status is bigger than the

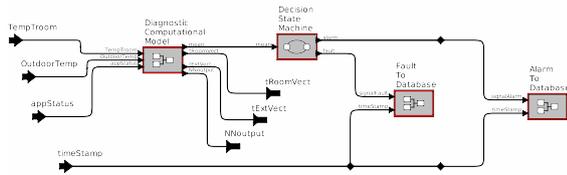


Fig. 5. PTOLEMY II model to detect air filter obstruction and generate corresponding alarms and faults.

corresponding fence value, a thermostat stuck-at fault is recorded.

As an example, in Figure 5 the PTOLEMY II model related to the air filter obstruction is shown. The main model components are DiagnosticComputationalModel, DecisionStateMachine, AlarmToDatabase and FaultToDatabase subsystems — data collection is not detailed in the figure. Bold arrows on the left of Figure 5 represent incoming data. DiagnosticComputationalModel subsystem contains actors capable of organizing raw data in vectors to be fed to an ANN to estimate the percentage of air coming through the warmer (0% fully obstructed - 100% no obstruction). The moving average of the estimated value is used as input of DecisionStateMachine subsystem where the proper event (i.e. *fault* or *alarm*) is determined. The decision is based on two thresholds $t_1 = \sigma$ and $t_2 = 2 * \sigma$ where σ is the standard deviation of the estimated percentage flow in normal conditions. If an alarm or a fault is detected, the corresponding event is inserted in the database by AlarmToDatabase and FaultToDatabase subsystems. In the plot of Figure 6 an example behavior of the HVAC system leading to identification of an air filter obstruction is shown. To generate the profile shown in the figure, a fault is injected into a simulated HVAC filter for a specified time interval. The onset and the end of the faulty behavior are marked by arrows in the plot. The profile of the fault is assumed in this case to be trapezoidal, i.e., starting with no obstruction the air flow is gradually reduced to 70% of the capacity and then it is gradually restored. The behavior of Figure 6 corresponds to the generation of several alarm events as soon as threshold t_1 is exceeded due to the initial drift with respect to the normal behavior, and then fault events when threshold t_2 is exceeded due to persistent anomalous behavior.

D. Web Services

Data coming from physical systems are collected in *xml* files and sent to the DDSS input web service through the Internet. Because of potential security threats, files are digitally signed combining a message-digest algorithm with public-key cryptography. Encryption uses the symmetric-key algorithm available in the Java security API. The code that implements web services consists of a manually-developed skeleton — which is invariant across applications — and application-dependent metadata. These are stored in tables inside DiSeGnO

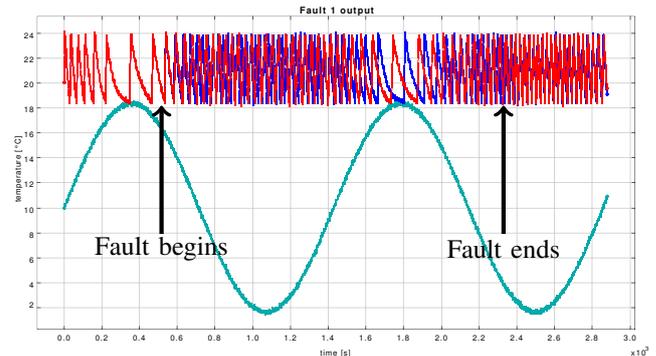


Fig. 6. An example of air filter obstruction. The x axis reports time (seconds) and the y axis reports temperature (Celsius degrees). The blue profile is the normal behavior, whereas the red profile is obtained by injecting the fault in the system.

TABLE I
LOAD RESULTS OF THE DDSS FOR DIFFERENT CONFIGURATIONS.

#	Hits	Mean service time [s]	Mean SQL time [s]	SQL hits/min	HTTP hits/min	Mean client time [h:m:s]
5	3600	2.611	0.597	236712	101	0:33:40
10	7200	4.073	0.871	264612	113	1:02:01
15	10800	6.557	1.330	309150	132	1:19:31
20	14400	9.070	1.774	302727	129	1:49:49
25	18000	10.894	2.123	315289	134	2:11:54

data store and contain all the information related to the queries that input and output web services have to execute. Metadata are leveraged by the skeleton to implement domain-specific behavior.

V. EXPERIMENTAL EVALUATION

The purpose of our experiments is to assess whether the DiSeGnO framework can be used in a real case, both in terms of absolute performances, i.e., time to store data coming from clients, and in terms of scalability, i.e., growth of computation time related to the quantity of data to be handled by the rules. Synthetic data are generated by PTOLEMY II models of the HVAC system shown in Figure 2. Data from temperature sensors and warmer apparatus are sampled (1 sample per simulation minute), collected in a file, and sent to the DDSS input web-services through an HTTP connection every 60 simulation minutes. Experiments were performed on a family of six identical Intel-based PCs, featuring a Core2Duo 2.13 GHz CPU, 4 GB of RAM and running Ubuntu Linux 12.04 (64 bit edition). Five “client” PCs run household simulations, and one “server” runs the DDSS server generated by DiSeGnO. Each client PC simulates 1 to 5 HVAC systems running for 30 simulated days. Server performances are monitored using JavaMelody², an open-source tool to profile Java server applications.

Table I shows load results for different configurations obtained varying the number of HVAC systems (“#”)

²<https://code.google.com/p/javamelody>.

TABLE II
REAL-TIME PERFORMANCES OF THE DDSS.

Number of rules	Wall clock time [h:m:s]	User time [s]	System time [s]
1	6:29:32	10539	1665
2	9:56:07	24340	1652
3	12:06:08	31600	1628

connected to the DDSS server on a time span of 30 days. For each row, the table shows the number of files sent to the DDSS server (“Hits”), the mean time to serve each file (“Mean service time”), the mean time to execute SQL queries related to a single file (“Mean SQL time”), the number of SQL queries per minute (“SQL hits/min”), the number of HTTP requests per minute (“HTTP hits/min”), and the mean time required by the client to send all the data (“Mean client time”). Notice that the figures for mean service time and mean database access time refer to cumulative performances averaged over the number of hits. On the other hand, mean client time refers to cumulative performances averaged over the number of systems. For instance, the last line of the table refers to loading data from 25 systems running for 30 (simulated) days. Since the simulation on the clients is accelerated, it takes only about two hours (on average) for a client to send all the data it generates in this case. Clearly, if the number of systems to monitor grows, the throughput of the DDSS decreases and client time increases — linearly in all the experiments we consider. However, two hours is about 2 orders of magnitude less than 30 days, indicating that the DDSS generated by DiSeGnO could support more systems or more signals easily. Table II shows the performances of the generated DDSS when varying the number of rules (1 to 3) applied to the biggest configuration loaded (25 systems running). Here, one can observe that the total wall clock time required on the server side is much less than 30 days, indicating that, even in its prototypical stage, the DDSS generated by DiSeGnO could run in real-time. On the other hand, the CPU time required to process the diagnostic rules (“User time” plus “System time”) albeit a fraction of total wall clock time — from 45% in the case of 1 rule, up to 73% in the case of three rules — indicates that the current implementation is apt to scale better in the number of systems rather than in the number of rules.

VI. CONCLUSIONS

Summing up, this paper shows that it is possible to combine ontology-based system descriptions and actor-based rule computation models in DiSeGnO framework to generate efficient DDSS software in a push-button way. In the current prototype implementation, DiSeGnO still relies heavily on PTOLEMY II to run the rules engine potentially requiring more computation time than an equivalent, manually-coded, DDSS. However, even in its present prototypical stage, the system is usable in practice to diagnose small-to-medium scale systems

with acceptable performances as shown in Table I and Table II. One of the issues left for future extensions is to automatically compile the model of rules in order to improve performances, e.g., by generating code independent from PTOLEMY II. A practical implementation on a real industrial case study will be the final testing ground.

REFERENCES

- [Agh85] G.A. Agha. *Actors: A Model Of Concurrent Computation In Distributed Systems*. PhD thesis, University of Michigan, 1985.
- [CGL⁺05] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. *DL-Lite: Tractable Description Logics for Ontologies*. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 602. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.
- [CM94] Sharma Chakravarthy and Deepak Mishra. Snoop: An expressive event specification language for active databases. *Data & Knowledge Engineering*, 14(1):1–26, 1994.
- [EJL⁺03] Johan Eker, Jorn W Janneck, Edward A. Lee, Jie Liu, Xiaojun Liu, Jozsef Ludvig, Sonia Sachs, Yuhong Xiong, and Stephen Neuendorffer. Taming heterogeneity - the Ptolemy approach. *Proceedings of the IEEE*, 91(1):127–144, 2003.
- [FMMV16] Francesco Flammini, Stefano Marrone, Nicola Mazzocca, and Valeria Vittorini. Fuzzy decision fusion and multiformalism modelling in physical security monitoring. In *Recent Advances in Computational Intelligence in Defense and Security*, pages 71–100. Springer, 2016.
- [GI13] Marco Gribaudo and Mauro Iacono. *Theory and Application of Multi-Formalism Modeling*. IGI Global, 2013.
- [GMF⁺03] J.H. Gennari, M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubézy, H. Eriksson, N.F. Noy, and S.W. Tu. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.
- [Gru95] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 43(5):907–928, 1995.
- [Hea14] Jeff Heaton. Encog machine learning framework, 2014. <https://github.com/encog>.
- [Kaz08] Y. Kazakov. *RIQ and SROIQ are Harder than SHOIQ*. In *Description Logics*, 2008.
- [KPJ06] Sanem Kabadayi, Adam Pridgen, and Christine Julien. Virtual sensors: Abstracting data from physical sensors. In *Proceedings of the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks*, pages 587–592. IEEE Computer Society, 2006.
- [LTSS11] Edward A Lee, Stavros Tripakis, Christos Stergiou, and Chris Shaver. A modular formal semantics for ptolemy. Technical report, University of California at Berkley — Dept. of Electrical Engineering and Computer Science., 2011.
- [MPSP⁺09] B. Motik, P.F. Patel-Schneider, B. Parsia, C. Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler, and et al. OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax. *W3C Recommendation*, 27, 2009.
- [NAL⁺07] SETU MADHAVI Namburu, Mohammad S Azam, Jianhui Luo, Kihoon Choi, and Krishna R Pattipati. Data-driven modeling, fault diagnosis and optimal sensor selection for hvac chillers. *Automation Science and Engineering, IEEE Transactions on*, 4(3):469–473, 2007.
- [RMC12] M. Rodriguez-Muro and D. Calvanese. Quest, an OWL 2 QL Reasoner for Ontology-based Data Access. *OWLED 2012*, 2012.
- [RWLF04] Kurt W Roth, Detlef Westphalen, Patricia Llana, and Michael Feng. The Energy Impact of Faults in US Commercial Buildings. In *International Refrigeration and Air Conditioning Conference*, 2004.

CHARACTERIZING WEB SESSIONS OF E-CUSTOMERS INTERESTED IN TRADITIONAL AND INNOVATIVE PRODUCTS

Grażyna Suchacka
Institute of Mathematics and
Informatics
Opole University
ul. Oleska 48
45-052 Opole, Poland
gsuchacka@math.uni.opole.pl

Grzegorz Chodak
Department of Operational
Research
Wroclaw University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
grzegorz.chodak@pwr.wroc.pl

KEYWORDS

Web traffic analysis, customer behavior, user session, click-stream analysis, log file analysis, Web server, e-commerce, innovative products

ABSTRACT

Web traffic characterization and modeling is currently a hot research issue. Low-level analysis of HTTP traffic on the server allows one to build adequate traffic models to be used in server benchmarking. High-level analysis of Web user behavior allows one to optimize website structure and develop personalized service strategies. In this paper, analysis of customer sessions in an online store is performed using Web server log data. The goal is to explore possible differences between sessions of customers viewing and purchasing innovative products, and customers only interested in traditional products.

INTRODUCTION

Electronic commerce has revolutionized the customers' approach to searching and buying products and services. Web users may access online stores regardless of time and space limits. A single visit of a customer to an online store corresponds to a series of Web pages opened by the customer via their Internet browser (a customer click-stream) and is called a *customer session* or a *Web session*.

The electronic environment makes it possible to collect detailed data on customers' behavior during their visit to an online store. This data may concern customers' searches, items viewed or added to the shopping cart, actions of purchase confirmation or withdrawal from the purchase, time spent in the store on performing various actions, etc. Basic data at HTTP level is recorded in standard Web server access log files. Reconstruction of customer sessions from logs is not a trivial task but it is worth an effort as it gives a researcher the ability to perform detailed analyses of many aspects of the Web traffic.

Various data mining methods have been applied so far to extract valuable knowledge on Web users' behavior and to predict customers' needs. The results have been used e.g., to improve customer relationship management (Xu and Wang 2011), to develop product

recommendation and search support systems (Cho et al. 2013), (Kuang and Li 2014), (Huk et al. 2015), to manage the store inventory (Chodak 2011), or to predict sales (Mohammadnezhad and Mahdavi 2012), (Suchacka and Chodak 2013).

A significant aspect of a click-stream analysis has been the discovery of user navigation patterns and developing models of user sessions (Krishnamurthy et al. 2010), (Kwan et al. 2005), (Nenava and Choudhary 2013), (Shim et al. 2012). Traffic models may be used to generate a synthetic click-stream at the server input to test the server performance under realistic and controllable workloads. The models may also be used to develop mechanisms aiming at improving the quality of Web service and predicting Web performance (Borzemski and Kamińska-Chuchmała 2012), (Borzemski and Suchacka 2010), (Zatwarnicki and Zatwarnicka 2014), (Zhou et al. 2006).

In this paper, we focus on the statistical analysis of sessions performed by customers interested in innovative products compared to sessions of customers only interested in traditional products available in an online store. The comparison is done in terms of such session characteristics as the session length, session duration, and mean time per page. Our motivation was to characterize the behavior of e-customers (especially buyers) interested in innovative products as a group which potentially copes better than other users in a virtual environment, which itself has already an innovative character. The results may be interesting and useful for an online store manager.

ANALYSIS OF CUSTOMER SESSIONS IN AN ONLINE STORE

In the e-commerce environment particular attention should be paid to the analysis of sessions ending with a purchase (i.e. *buying sessions*) as they provide the online retailer with very important information, related to revenues and profits. This information concerns not only products bought by customers, but also the earlier customers' behavior in the store and sources of their visits (a reference from an organic or paid search engine result, a reference from an e-mail newsletter, entrance through a social media website, etc.). From a store manager's point of view, the analysis of buying sessions has the following objectives:

- analysis of sources that referred customers to the store (this allows the retailer to assess the effectiveness of various marketing channels and their rates of return);
- customer segmentation in respect of further marketing activities (mailings, recommendation system, re-marketing activities, etc.);
- optimization of the store offers in terms of potential demand (e.g., based on information about categories of products viewed by customers and keywords typed into the store's search engine);
- optimization of the website's content, including the user interface;
- optimization of shipping and payment forms.

In practice, customers visiting online stores reveal differentiated navigational and behavioral patterns. This fact was a motivation to apply various segmentation or clustering techniques to determine customer profiles (Nenava and Choudhary 2013), (Song and Shepperd 2006), (Tanna and Ghodasara 2012), (Wang et al. 2004). In other studies VIP customers (Shim et al. 2012) or loyal customers (Chang et al. 2007) have been identified.

Since we wanted to investigate whether there are differences in traffic characteristics for customers interested in innovative and traditional products, we decided to distinguish between two subgroups of e-customers taking into account their preferences for types of viewed products. We have defined two types of customers in an online bookstore :

- 1) *Innovative customers (I)* are users who viewed some products considered "innovative", i.e. audio-books and multimedia products;
- 2) *Traditional customers (T)* are users who did not view any innovative products, but only traditional products, i.e. printed books.

RESEARCH METHODOLOGY

Raw data used for the analysis had been recorded from 1 April to 30 September 2014 in access logs of a Web server hosting the online bookstore (the address of the store website is not given in the paper due to a non-disclosure agreement). Data was written in the NCSA Combined log format. Each single HTTP request was described with the following data:

- IP address of the HTTP client (Web browser),
- date and time stamp meaning the time of the request coming to the server,
- HTTP method,
- version of HTTP protocol (1.0 or 1.1),
- HTTP status code,
- URI identifying the requested server resource,
- volume of data (in bytes) sent from the server in response to the request,
- URL of a site which linked the user to the store,
- user agent string, containing the name and version of the client Web browser.

A dedicated computer program was written in C++ to read, preprocess, clean, and analyze the data. Of all the HTTP requests we left only requests corresponding to

page views (user clicks), eliminating requests for embedded objects (graphics and video files, etc.). Based on page view requests, click sequences for individual Web users were identified. Each unique user was identified based on two request data fields combined: the IP address of the client and the user agent string. Afterwards, Web sessions for each user were identified assuming that a minimum interval between two subsequent sessions of a given user is 30 minutes. Since our goal was the analysis of sessions performed by customers viewing and buying products in the online store, we eliminated sessions issued by the website administrator and sessions performed by Internet bots. Bots were identified using a methodology proposed in (Suchacka 2014). We also eliminated sessions containing only one page view and/or lasting less than two seconds.

To verify and refine the description of customer sessions, we used data that had been gathered by the tracking software, SuperTracker, for the analyzed website. To obtain information on categories associated with products viewed in customer sessions, we combined data from three sources: server logs, SuperTracker database, and product database.

The entire data set contained 33 354 customer sessions. The subset of *innovative customer* sessions contained 6 171 sessions (including 466 buying ones) and the subset of *traditional customer* sessions contained 5 415 sessions (including 207 buying ones).

For each session three features were determined:

- session length – the number of pages opened during the session,
- session duration – the time interval between the last and the first customer's clicks in session,
- mean time per page – the average time of browsing a single page in session.

Such features of e-customers' visits may be obtained via some analytical tools, e.g., Google Analytics (GA). However, GA statistics may be less accurate than information gathered directly from logs because they are based on a sampling procedure (Google 2016). Session sampling may result in inaccurate results especially in the conversion analysis where conversions constitute a small fraction of all sessions. Furthermore, online retailers are not willing to share high-level information about traffic on their websites. Our software has the advantage that it uses low-level data written in server logs which are much easier to obtain.

RESULTS AND DISCUSSION

We analyzed session characteristics for *innovative* and *traditional* customers both for all sessions in each group and only for buying sessions in each group. The results are shown in Tab. 1-3.

Fig. 1 and Fig. 2 show histograms of session lengths, session durations, and mean times per page for all *innovative* and *traditional* customer sessions, respectively. To compare the results for both customer types, histogram intervals have a fixed width (the figures show up to fifteen first intervals).

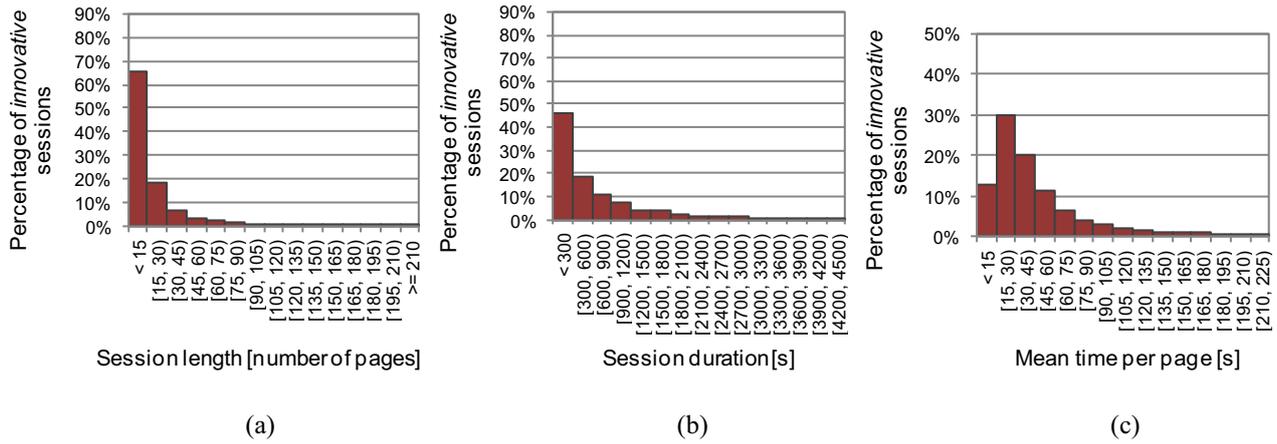


Figure 1: Data distributions for all *innovative customer* sessions

(a) Histogram of session lengths; (b) Histogram of session durations; (c) Histogram of mean times per page

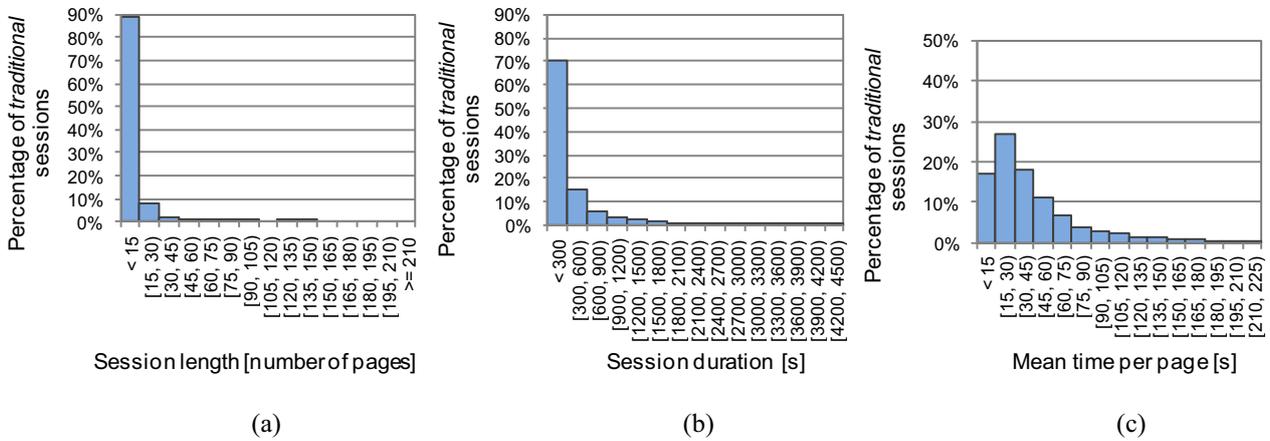


Figure 2: Data distributions for all *traditional customer* sessions

(a) Histogram of session lengths; (b) Histogram of session durations; (c) Histogram of mean times per page

Table 1: Session Length Statistics (in Number of Pages in Session)

Statistics	All sessions		Buying sessions	
	<i>I</i>	<i>T</i>	<i>I</i>	<i>T</i>
Mean	17.4	7.7	60.9	37.3
Median	9	5	52	33
Mode	3	3	31	23
Std. dev.	23.7	9.3	36.9	19.2
Minimum	2	2	9	13
Maximum	286	145	264	145

Table 2: Session Duration Statistics (in Seconds)

Statistics	All sessions		Buying sessions	
	<i>I</i>	<i>T</i>	<i>I</i>	<i>T</i>
Mean	680	327	1 801	1 164
Median	344	147	1 476	859
Mode	39	23	1 290	618
Std. dev.	893	491	1 337	926
Minimum	2	2	195	142
Maximum	10 043	7 544	10 043	5 772

Table 3: Mean time per page statistics (in Seconds)

Statistics	All sessions		Buying sessions	
	<i>I</i>	<i>T</i>	<i>I</i>	<i>T</i>
Mean	69.4	67.7	31.8	34.1
Median	34.6	34.0	27.1	27.5
Mode	15.0	21.0	37.6	36.4
Std. dev.	131.2	126.6	19.0	25.4
Minimum	0.5	0.5	5.9	6.6
Maximum	1 788.0	1 641.0	171.0	197.6

One can notice in Tab. 1 that visitors interested in traditional, printed books typically open two times less pages in a session (7.7 on average) than customers searching for audio-books (17.4 on average). For *innovative customer* sessions both the mean and the median are much higher and their session lengths are more differentiated (as indicated by very high value of the standard deviation). For both customer groups histograms of session lengths (Fig. 1a, Fig. 2a) are right-skewed and long-tailed – very short sessions dominate especially in the case of *traditional customers*.

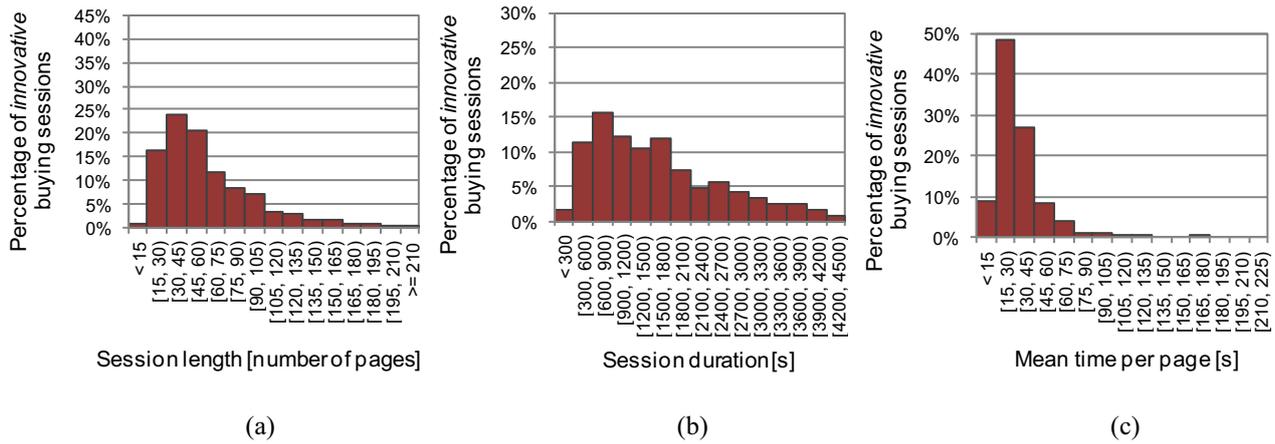


Figure 3: Data distributions for buying *innovative customer* sessions

(a) Histogram of session lengths; (b) Histogram of session durations; (c) Histogram of mean times per page

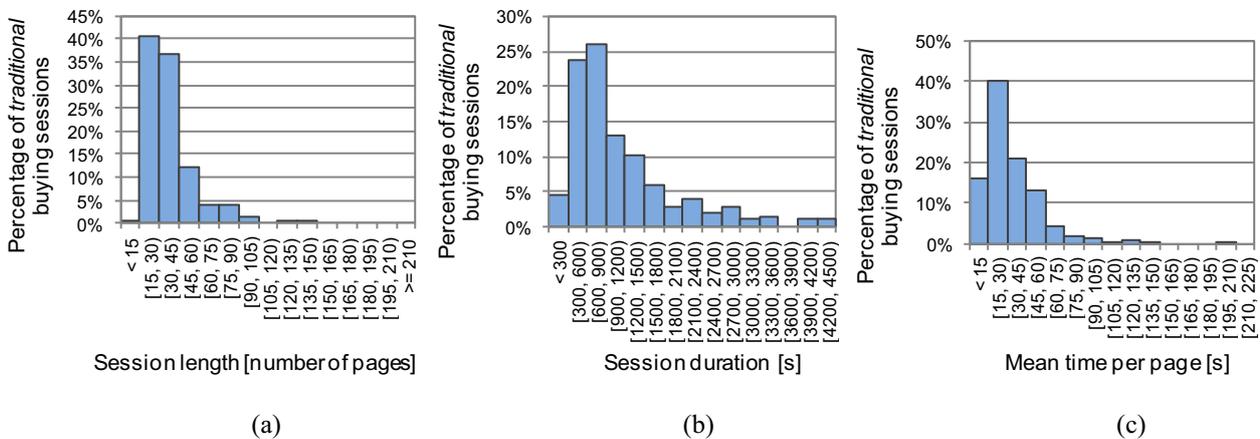


Figure 4: Data distributions for buying *traditional customer* sessions

(a) Histogram of session lengths; (b) Histogram of session durations; (c) Histogram of mean times per page

Session lengths correspond to session durations (Tab. 2, Fig. 1b, Fig. 2b). *Innovative customers* spend on average much more time interacting with the site (11.3 minutes) than *traditional* ones (5.4 minutes). However, statistics for the mean time per page (Tab. 3) are very similar for both groups – the mean is equal to about 1 minute (69.4 seconds for *innovative customers* and 67.7 for *traditional* ones). Such results suggest that generally customers do not differ much in the way of browsing pages and analyzing Web store content; however, they may have differentiated needs and expectations which cause *innovative customers* to perform more searching and browsing operations in the bookstore.

Differences in session characteristics for both customer groups are even more visible in the case of buying sessions (c.f. statistics for buying sessions in Tab. 1-3). Moreover, for each group the buying session characteristics differ from the general results in each group: buyers open many more pages (*innovative*: 60.9 pages compared to 17.4 pages, *traditional*: 37.3 pages compared to 7.7 pages) and spend much more time on the site (*innovative*: 30 minutes compared to 11.3 minutes, *traditional*: 19.4 minutes compared to 5.4

minutes), whereas the mean time per page is much shorter (more than half a minute compared to more than one minute for both customer groups). Buyers seem to only recently be familiar with the offers of the online store and do not browse information pages so intensively (moreover, the checkout process itself includes several pages informing them about the ordered products, the total cost, conditions and address of delivery, etc., and interaction with these sites is usually not very time-consuming, especially for regular customers).

For buyers, distributions of session lengths (Fig. 3a, Fig. 4a), session durations (Fig. 3b, Fig. 4b) and mean times per page (Fig. 3c, Fig. 4c) differ between the customer groups. In the case of *traditional* buyers short sessions (containing 15 – 45 pages, lasting 5 – 15 minutes) clearly dominate; furthermore, the session length (Fig. 4a) and session duration (Fig. 4b) distributions are clearly right-skewed, whereas the shapes of these histograms for *innovative* buyers (Fig. 3a, Fig. 3b) are different. As regards the mean time per page, it tends to be higher for *innovative* buyers than for *traditional* ones (Fig. 3c, Fig. 4c).

The differences between the statistics for two groups of buyers may be caused by the innovative products' descriptions which often include multimedia samples or film presentations (trailers, etc.). The multimedia content located on the product description page usually requires the customer to devote more time to them than other product description pages. Another hypothetical explanation may be such that *innovative customers* are more computer-oriented and therefore they spend more time in the online store as they feel more comfortable in a digital online store than *traditional customers*.

CONCLUSIONS

The paper discusses results of the statistical analysis of sessions performed by customers viewing and buying innovative and traditional products in an online bookstore. The results show that customers viewing and buying innovative products open on average many more pages and spend much more time interacting with the site than in the case of traditional products. This tendency is even more visible if we confine ourselves only to buying sessions.

The resulting distributions of the session lengths, session durations, and mean times per page may be used in simulation models. They may also be useful in setting input values to other data mining methods, e.g. association rules, applied to online store data. In a broader perspective the results may be used to personalize and improve the future customers' online shopping experience. It should be kept in mind, however, that user visits to different online stores may be characterized by different navigation patterns and thus, the results of our analysis cannot be automatically generalized to other e-stores.

In future work we are planning to extend our research to datasets from other e-commerce sites. Furthermore, as we considered a very limited number of session features in this study, we are planning to investigate intra-session dependencies between other features of sessions performed by customers interested in innovative products. Finally, it would be very interesting to apply some unsupervised machine learning methods to determine customer groups based on product categories rather than relying on arbitrary defining two customer subgroups. In this respect, we are planning to apply clustering methods to automatically divide e-customers into clusters using information on product categories.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Centre (NCN) in Poland under grant no. 2013/11/B/HS4/01061.

REFERENCES

Borzemski, L. and A. Kamińska-Chuchmała. 2012. "Client-perceived Web performance knowledge discovery through turning bands method." *Cybernetics and Systems* 43, No.4, 354-368.

- Borzemski, L. and G. Suchacka. 2010. "Discovering and usage of customer knowledge in QoS mechanism for B2C Web server systems." In *Proceedings of the 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Cardiff, Wales, Sep.08-10), Lecture Notes in Artificial Intelligence 6277, Part II. Springer, Berlin Heidelberg, 505-514.
- Chang, H.-J.; L.-P. Hung; and C.-L. Ho. 2007. "An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis." *Expert Systems with Applications* 32, No.3, 753-764.
- Cho, Y.S.; S.C. Moon; S.-p. Jeong; I.-B. Oh; and K.H. Ryu. 2013. "Clustering method using item preference based on RFM for recommendation system in u-commerce." *Lecture Notes in Electrical Engineering* 214, *Ubiquitous Information Technologies and Applications*, 353-362.
- Chodak, G. 2011. "ABC analysis in an internet shop: a new set of criteria." In *Proceedings of the IADIS International Conference* (Avila, Spain, Mar.10-13), 196-204.
- Google. 2016. "How sampling works." Analytics Help, <https://support.google.com/analytics/answer/2637192?hl=en> (access: Feb.11).
- Huk, M.; J. Kwiatkowski; D. Konieczny; M. Kędziora; and J. Mizera-Pietraszko. 2015. "Context-sensitive text mining with fitness leveling genetic algorithm." In *Proceedings of the 2nd International Conference on Cybernetics CYBCONF'15* (Gdynia, Poland, Jun.24-26). IEEE, New York, USA, 183-188.
- Krishnamurthy, D.; M. Shams; and B.H. Far. 2010. "A model-based performance testing toolset for Web applications." *Engineering Letters* 18, No.2, 92-106.
- Kuang, G. and Y. Li. 2014. "Using fuzzy association rules to design e-commerce personalized recommendation system." *TELKOMNIKA Indonesian Journal of Electrical Engineering* 12, No.2 (Feb), 1519-1527.
- Kwan, I.S.Y.; J. Fong; and H.K. Wong. 2005. "An e-customer behavior model with online analytical mining for internet marketing planning." *Decision Support Systems* 41, No.1, 189-204.
- Mohammadnezhad, M. and M. Mahdavi. 2012. "Providing a model for predicting tour sale in mobile e-tourism recommender systems." *International Journal of Information Technology Convergence and Services (IJITCS)* 2, No.1 (Feb), 1-8.
- Nenava, S. and V. Choudhary. 2013. "Hybrid personalized recommendation approach for improving mobile e-commerce." *International Journal of Computer Science & Engineering Technology (IJCSSET)* 4, No.5, 546-552.
- Shim, B.; K. Choi; and Y. Suh. 2012. "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns." *Expert Systems with Applications* 39, No.9, 7736-7742.
- Song, Q. and M. Shepperd. 2006. "Mining Web browsing patterns for e-commerce." *Computers in Industry* 57, No.7, 622-630.
- Suchacka, G. and G. Chodak. 2013. "Practical aspects of log file analysis for e-commerce". In *Proceedings of the 20th International Conference Computer Networks CN'13* (Lwówek Śląski, Poland, Jun.17-21), Communications in Computer and Information Science 370. Springer, Berlin Heidelberg, 562-572.
- Suchacka, G. 2014. "Analysis of aggregated bot and human traffic on e-commerce site." In *Proceedings of Federated Conference on Computer Science and Information Systems FedCSIS'14* (Warsaw, Poland, Sep.7-10), ACSIS, Vol. 2. IEEE, New York, USA, 1123-1130.

- Tanna, P. and Y. Ghodasara. 2012. "Exploring the pattern of customer purchase with Web usage mining." In *Proceedings of the International Conference on Advances in Computing ICAdC'12* (Bangalore, India, Jul.4-6), AISC 174. Springer India, New Delhi, India, 935-941.
- Wang, Q.; D.J. Makaroff; and H.K. Edwards. 2004. "Characterizing customer groups for an e-commerce website." In *Proceeding of the 5th ACM Conference on Electronic Commerce*. ACM Press, New York, 218-227.
- Xu, H. and L. Wang. 2011. "Application of analysis CRM based on association rules mining in variable precision rough set". In *Proceedings of International Conference on Computer Science, Environment and Ecoinformatics* (Wuhan, China, Aug.21-22), Communications in Computer and Information Science 216. Springer, Berlin Heidelberg, 418-423.
- Zatwarnicki, K. and A. Zatwarnicka. 2014. "The cluster-based time-aware Web system." In *Proceedings of the 21th International Conference Computer Networks CN'14* (Lwówek Śląski, Poland, Jun.23-27), Communications in Computer and Information Science 431. Springer, Berlin Heidelberg, 37-46.
- Zhou, X.; J. Wei; and C.-Z. Xu. 2006. "Resource allocation for session-based two-dimensional service differentiation on e-commerce servers." *IEEE Transactions on Parallel and Distributed Systems* 17, No.8, 838-850.

AUTHOR BIOGRAPHIES

GRAŻYNA SUCHACKA received the MS degrees in Computer Science and in Management from Wrocław University of Technology, Poland. She received her Ph.D. degree in Computer Science from Wrocław University of Technology. Now she is an assistant professor in the Institute of Mathematics and Informatics at Opole University, Poland. Her research interests include Web mining, Web analytics, and Quality of Web Service with special regard to electronic commerce. Her e-mail address is: gsuchacka@math.uni.opole.pl.

GRZEGORZ CHODAK received the MS degree in Computer Science from Wrocław University of Technology, Poland. He received his Ph.D. degree and habilitation in Management from Wrocław University of Technology. Now he is an assistant professor in the Department of Operational Research at Wrocław University of Technology. His research interests include e-commerce, logistics with regard to online stores. His e-mail address is: grzegorz.chodak@pwr.wroc.pl.

CLOUD IMPLEMENTATION OF AGENT-BASED SIMULATION MODEL IN EVACUATION SCENARIOS

Andrzej Wilczyński
Cracow University of Technology
Warszawska 24, 31-155 Cracow, Poland
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Cracow, Poland
E-mail: and.wilczynski@gmail.com

Joanna Kołodziej
Cracow University of Technology
Warszawska 24, 31-155 Cracow, Poland
E-mail: jokolodziej@pk.edu.pl

KEYWORDS

Computational Cloud; Multi-Agent System; Evacuation System; Modelling and Simulation; MapReduce; Hadoop.

ABSTRACT

Over the years evacuation simulation has become increasingly important in the research on the wide class of problems related to the public security in emergency situations. In this paper we develop simulation platform fully integrated with the cloud system with using the MapReduce programming model and Hadoop framework. The environment illustrating evacuation scenarios and actors is modelled, by cell automata and interpreted as a potential field, in which technologies the generated agents are located with using multi-agent. The simulation is executed as a stream-based data-processing operation to enable environmental universality while taking advantage of the MapReduce model. Several test cases are provided to show the efficiency of the simulation platform.

INTRODUCTION

The management of emergency activities such as guiding people out of dangerous areas and coordinating rescue teams is characterized by uncertainty regarding both the source of danger and the availability of useful resources. Depending upon the scale and nature of the incident, people involved in a crisis may suffer from limited situational awareness (SA) [34]. SA involves being aware of what is happening around in order to understand how information, events, and the crowd actions will impact the goals and objectives.

Lacking or inadequate SA in emergency situations has been identified as one of the primary factors leading to human error, with potentially grave consequences. However, emergency situations are chaotic in nature and any incident management system typically encompasses multiple sources of information such as mobile devices from affected people, social network feeds and various on-site sensors. All of these sources are available in different formats, present in different locations and reliable at different confidence levels. As sources are dispersed throughout geographical areas, SA becomes a complex, distributed processing problem

where innovative techniques need to be employed in order to efficiently use information sources in real-time. In many cases it is impossible to provide an effective crowd management in evacuation by extracting data generated in real-world evacuation [1, 2] and realistic experiments [3, 4], mainly because of large data volumes and incompleteness. In such cases, simulation can be the proper solution for such problems.

Modeling and simulation methods designed for evacuation systems can be divided into two main categories: (i) flow-based approaches and (ii) individual-based approaches. In flow-based approaches [5] the behavior of individuals in the crowd is ignored, while in the second category [6] the crowd is defined as a collection of individuals, which are either entity-based or agent-based with autonomous intelligent agents.

Agent-based approaches typically focus on defining the rule of an individual's behavior and then apply the rule to all individuals of the whole simulated crowd [7-11]. However, the complexity of MAS in big crowd simulations is high, which makes the whole model ineffective in the case of short-time decision processes such as crowd management in emergency scenarios. Therefore, typical MAS used for crowd modeling has been supported additionally by using various multi-CPU systems [12], cluster [13] grid technologies [14-16], Graphic Processing Unit (GPUs) [17] and Field Programmable Gate Array (FPGAs) [18].

Cloud computing entails the exchange of computer and data resources across global networks; it constitutes a new value-added paradigm for network computing, where higher efficiency, massive scalability and speed rely on effective software development [19-20]. Cloud computing is rapidly becoming a popular infrastructure of choice among all types of organizations. Despite some initial security concerns and technical issues, an increasing number of institutions are considering moving their applications and services into "The Cloud." Recently, cloud-based technologies have been successfully used as effective support tools in complex simulations in emergency situations [21-23]. MapReduce programming model [24-26] is very popular tool for implementation of the crowd simulation in evacuation situations as two-stage *mapper-reducer* process. In *mapper* phase, the input data is extended

into a large intermediate table, while in the *reducer* phase table is reduced in order to generate the output data. The MapReduce procedure can be iterated many times until specified stopping criteria (usually the expected output data).

There are not many successful examples of integration of evacuation MAS with cloud environments. In most of such approaches, MAS-based simulation is interpreted as a simple data processing operation. In [27], in the loop-based simulation process, agents have been ordered and processed as a queue in the loop. The agent queues have been implemented as big tables of various sizes generated during the simulation. In this paper, we develop OpenStack [28] cloud-based simulation platform using the MapReduce programming model to support a large-scale evacuation crowd simulation. Multi-agent technology was employed to provide the crowd simulation and a grid-based model was used to provide environmental information. The simulation results show benefits of using the cloud-support in evacuation in restricted indoor environments compared to traditional IT support used in many realistic scenarios.

The paper is organized as follows first we provide a short description of the environmental model and cloud-based simulation platform. Next, we define the MapReduce-based multi-agent simulation model and specify the output data requests and generation. Then we present the results of simple experiment. After that we define draft conclusions and plans for future work in this domain.

OPEN STACK CLOUD-BASED SIMULATION PLATFORM FOR CROWD EVACUATIONS

We designed and implemented our cloud-based simulation platform by using the cloud OpenStack technology [28] and MapReduce programming model [24-26] with Hadoop framework [29]. The platform model architecture is presented in Fig. 1. The model is composed of three main modules, namely (i) a generic environment model, (ii) a multi-agent based simulation module and a data requestor.

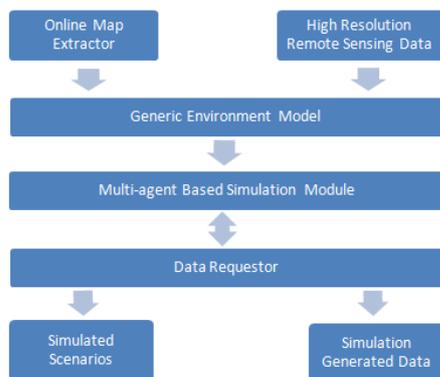


Figure 1. The Architecture of the Simulation Platform

The generic environment model is a formal model of the virtual environment used for simulation of the crowd evacuation scenarios. This model is based on the potential field paradigm and contains a vector field-based pathfinding mechanism. In the environmental model, various types of spatial data from multiple data sources (e.g., high-resolution remote sensing data or online map services data) are collected, merged and converted into streams and multi-streams for the further.

The multi-agent simulation module is used for simulations of the crowd behavior and decisions in emergency situations. It is based on the MapReduce architecture and runs on a Hadoop cluster. Most of the agent's operations are transformed into simple data access and processing operations. Hadoop Distributed File System (HDFS) is a standard technology used for data and information transfers.

GENERIC ENVIRONMENTAL MODEL

The generic environment model in this approach is based on environment models with cell automata developed in [30, 31]. The main component of this model is an *environmental map* defined as a grid in which each cell represents a small square area in physical environment. We specify three types of cells, namely: (i) accessible cells, (ii) blocked cells and (iii) exit cells. An accessible cell represents an area which may be accessed by the individual (human, agent), where blocked cell is defined as an area with obstacles. In the simulation, individuals are able to pass through accessible cells but not blocked cells. The environmental map can be generated from satellite photos (sensing data) or online web maps (e.g., Google Map).

The second component of the environmental model is a position potential field [32] for the management of the agent's position on the environmental map. A potential field module calculates a distance to the nearest exit cell from the specified (current) agent's position. Then it is responsible for the generation of the optimal path for the agent from his current position to the exit.

The position potential field is usually defined as a discrete 3D function with values numerically represented as a matrix of the same size as environmental map. Each parameter in that matrix is interpreted as the cell's 'position potential'. In particular, the position potential of obstacle cells is -2, which means that the corresponding cells are inaccessible. Obviously, all exit cells have a position potential of 0, which is the lowest valid value.

With the potential field, the general environment model is able to provide path information for the simulation, which can eliminate the need for pathfinding algorithms. In fact, in our agent implementation, routing

is only a search operation in a small list, which can be much faster than ordinary pathfinding algorithms are.

MAP-REDUCE BASED SIMULATION

The crowd behavior in evacuation scenarios is modelled by the multi-agent system (MAS). MAS in our approach is fully integrated with the SaaS (Software-as-a-Service) cloud layer and MapReduce model. The architecture of the multi-agent simulation module is shown in Fig. 2.

Crowd simulation is performed continuously as a phase loops (steps). Each loop corresponds to one time-step and contains one map operation and one reduce operation. After each loop, a position (location) of each agent is updated. It means that a serial cell selection operation is performed and the simulated scene of this time-step is produced.

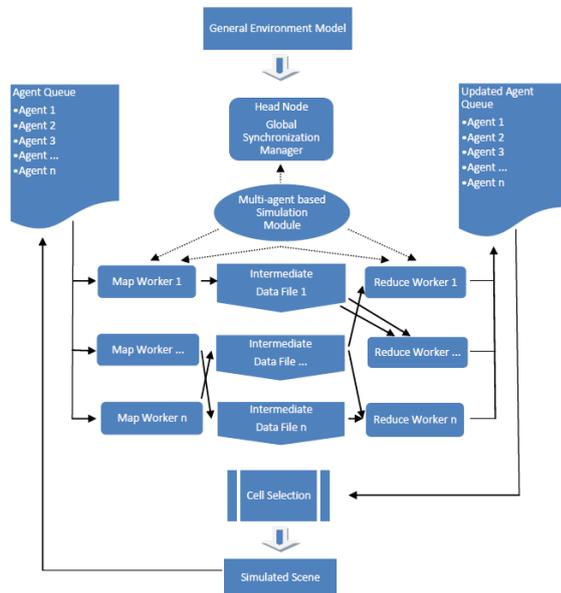


Figure 2. MapReduce-based Simulation Module Architecture

Agent Modelling

An agent contains the following information about its location, speed of movement, health, recent moves and a set of weights as well as a candidate cell queue. Cell queue consists of eight cells from the neighbourhood specified by the current agent's position in the environment. This situation is illustrated in Fig. 3. Each agent decides about possible movements in a given time step. If the agent decides to move, it can only move to one of the cells in the candidate list. More information about the agent model and its uses in the clouds is discussed in [27, 33].

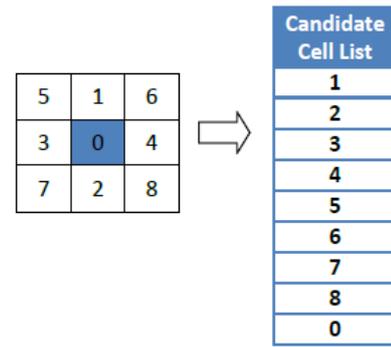


Figure 3. The Candidate Cell List

The crowd simulation process begins with the transformation of the agents in the agent items that contain key-value pairs, as shown in Fig. 4. A key is used to determine the position of the agent. It is unique due to the fact that in one cell it can be only one agent and it is static because the position does not change during the processing stage of the loop. An important element which deserves attention is that Hadoop sorts the agent items depending on their state of health. Then, agent items are transferred to the mapper which is described below.

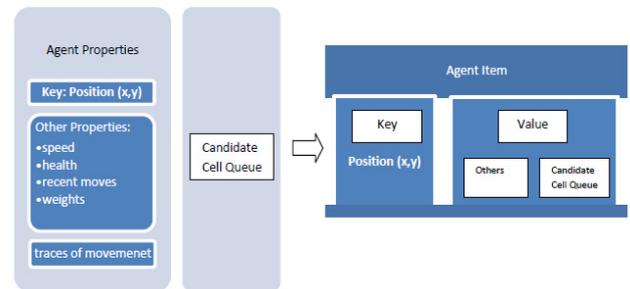


Figure 4. The Transition From Agent Model to Agent Item

The Mapper

The input of the agent is transmitted to the mapper that creates a set of intermediate key/value pairs, each of these pairs determines the attraction value of each item in the queue of agent's candidate cells. Identifier of each of these pairs is $iKey$, which means that the items in the queued candidate cells of the same agent have the same $iKey$. This phenomenon is characteristic of the reducer by the Hadoop framework. Environment data, which requires the mapper can be presented in the form of a general environment model. For small and simple problems, they can be transferred from the standard input or to speed up the simulation built into the mapper.

The mapper calculates the cell attraction value of all eight surrounding cells and produces an intermediate table, as shown in Fig. 5. The intermediate table contains information of the agent itself and a candidate cell queue with each cell's attraction value. The first

item corresponds to the current cell itself, which has the lowest attraction value since the agent on current cell will choose to stand still only if it cannot move to any of the 8 surrounding cells. Next, the intermediate table is sorted by Hadoop framework in descending order according to cell attraction value of each pair.

The Reducer

Agent items with the same *iKey* are combined to reproduce the agent information, as shown in Fig. 5. It is worth noting that the output of the reducer is the same as the input of the mapper, as shown in Fig. 6. Thus, the role of the reducer is to collect information from the sorted intermediate table and generate agent information, which is then transmitted to the Hadoop framework to create a list of agents.

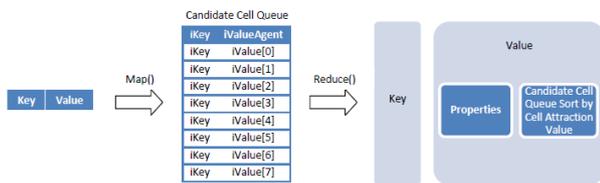


Figure 5. The Map-and-Reduce Process used in the Simulation

Cell Selection and Agent Movement

Every moving of an agent can affect other agents and attempt to parallel shift can cause errors. To prevent this happening, the parallel map and reduce operations only generates a sorted candidate cell queue for each agent but do not move agents. These movements are performed in a processing stage called a cell selection. Here, the agent queue is processed one by one. For each agent one cell is selected as a target, but it must be a cell that has not been previously taken by another agent. Then the agent is moved and the information about its location is updated.

After the cell selection stage, the simulated scene of the current time-step is fully ready. A new agent queue is also produced, which is the input for the next time-step.

Simulation Process Control and Data Generation

The simulation process is a process that takes place in an infinite loop, each iteration represents exactly one time-step, as shown in Fig. 6. It is an asynchronous process which is run in the cloud. However, if a user sends a request, the data requestor will send an imperative synchronization signal to the simulation module. At the time of request processing a global lock is set, which means that all nodes in the cloud are stop for a specified period of time. Then, the entire static (frozen) simulation scenario is generated by all the nodes and then returned to the user. After this process is completed, the global lock is canceled and nodes return to their previous jobs.

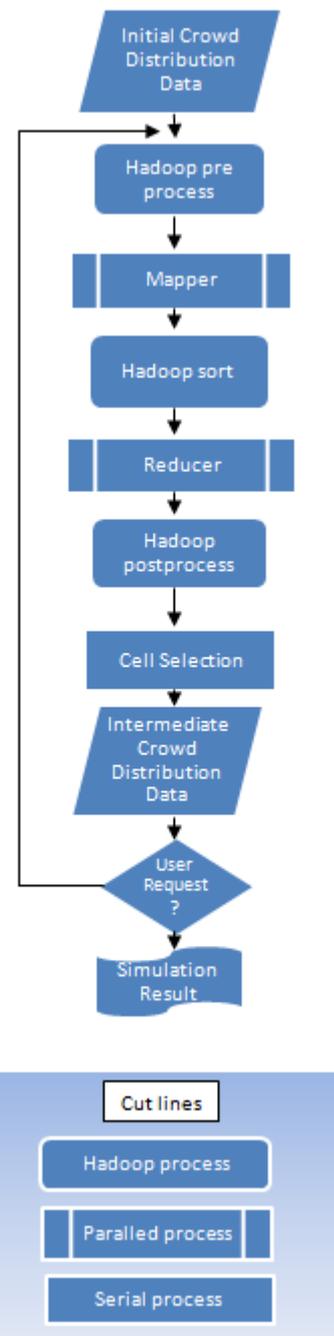


Figure 6. Simulation Process

CASE STUDY

In this Section we demonstrate results of simple experiments provided for verification of the correctness of platform implementation and cloud system motivation, and agent-based crowd management in the evacuation scenarios.



Figure 7. Environmental map model

The environmental model includes a simple map of 9000 m² indoor area with 4 exits is presented in Fig. 7. The tests were run on a small physical Hadoop cluster with one head node and 20 slave nodes with identical hardware and software configuration. This small physical cluster is connected with the wide public OpenStack cloud environment to provide proper services for crowd management. The mapper and the reducer module were written in C++ and implemented by using the Hadoop framework (streaming module). The environment information is generated by a map editor and compiled in both mapper and reducer modules. A platform evaluation process is monitored by the internal shell mechanisms in the cloud. The monitoring results are stored in the shellsripts. Those scripts are used also for activating the Hadoop framework and running the mapper and the reducer.

We compared the results achieved by the cloud-based Hadoop cluster with the result of similar tests performed for a single powerful workstation with the same and 10 times smaller crowd, where all operations were sequential without the cloud support. This scenario is very typical for conventional evacuation systems and most of realistic evacuation scenarios so far. The main idea of such a comparative analysis was to demonstrate benefits of using Hadoop clusters, even if the synchronization of parallel processes in the system may delay the whole crowd management. Table 1 shows hardware and software settings for Hadoop cluster and the sequential server.

TABLE I. CONFIGURATION OF HADOOP NODES AND THE POWERFUL SEQUENTIAL SERVER USED IN EXPERIMENTS

Hardware for the Server and a Hadoop nodes	CPU	Intel Core i7 3770 4x 3.4GHz
	RAM	8GB
	Network	1000Mbit Ethernet
Software	OS for Server	Windows Server 2010 64-bit
	OS Hadoop	Ubuntu 14.04 LTS 64-bit, hadoop 2.7.1

In our experiments, we first compare the simulation time of the crowd. We assume that in both systems –

cloud with Hadoop and sequential server – we have 20000 agents and the size of the crowd increases in equal time intervals by 20% of the maximal amount of agents – it means 4000 in each time interval. Another criterion was a simulation cost measured for in the sense of total memory usage and memory working set size recorded after 50 time intervals. The length of the one time interval in simulation was set to 3 sec. For this experiment we have generated 2 environmental scenarios: (i) for the Hadoop cluster, we generated a large number (20,000) of individuals (agents); (ii) for powerful sequential server we generated a smaller number (2,000) of individuals (agents) trying to evacuate themselves from the same building. The number of agents for servers represented 1/10 of the number of agents for Hadoop cluster with 20 nodes, which makes the tests comparable: the average number of agents which can be managed by a Hadoop node is 2000.

Crowd evacuation times

Table II gives the execution times (in seconds) of the two types of crowd simulation systems with increasing crowd sizes.

TABLE II SIMULATION TIME

Nb of agents	Server	Hadoop
4000	899	1429
8000	1990	1599
12000	2976	2009
16000	N/A	2198
20000	N/A	2244

In the case of 4000 agents, the execution time in cloud simulation is longer (almost 2 times) than in the case of sequential server. The reason can be the communication overhead of the cloud services. Hadoop becomes to work better for big crowds (over 10000 agents). In those scenarios, the communication overhead is covered by the benefits gained by distributing computation load of agents to multiple compute nodes. The achieved results indicate that (i) the cloud architecture can significantly improve the runtime performance of executing complex

simulation scenarios and (ii) it scales well with the scenario's complexity.

Simulation cost

Table III shows the calculated simulation costs after 50 time intervals including total memory usage, memory working set size and average memory usage for one individual for both Hadoop cluster and the sequential server.

TABLE III SIMULATION COST

	Hadoop	Server
Total memory usage (KB)	1056597	122632
Memory working set size (KB)	52113	18872
Memory usage per individual (KB)	65.660	61.316
Memory working set size per individual (KB)	3.211	9.436

It could be expected that the total memory utilization and the memory working set size are larger for the Hadoop cluster. This is mainly because of the initial population of the agents, which was 10 times bigger than that generated for the single server. However, memory distribution per individual is almost the same, which shows good memory management results in the cloud system. But the most significant results are for the fourth parameter in Table III: the memory working set size for each individual. In the case of Hadoop, this parameter is only 34% of the value achieved for the single server. The main reason is that in the cloud environment, data processing operations are processes successfully executed for different data sets. In the case of the single server, each data mining procedure is executed as a new process. This comparison shows significant potential benefits of using the cloud system for supporting the crowd evacuation in critical scenarios. Such benefits can be observed in the case of a big disproportion in the crowd size (10 times larger in the cloud-support case).

CONCLUSIONS AND FUTURE WORK

In this paper, we presented an early-stage development results on OpenStack cloud-based multi-agent simulation platform for evacuation of the crowd from the indoor environment with the limited number of evacuation exits and evacuation path size. Environment in this model is represented by cell automata and interpreted as a potential field, in which generated agents are located. The crowd management in the cloud is supported by the MapReduce programming model with the classical Hadoop framework used for its implementation. Simple experiments were performed on a small Hadoop cluster with ten nodes and separately for a single powerful server in order to demonstrate potential benefits of using the cloud system. The results of the experiments show that cloud-based systems can reduce significantly the complexity of the management of individuals in the crowd. Moreover, there is no need to initiate the large number of new processes on the

same work station cause some data processing operations can be performed by using the software frameworks shared inside the public cloud.

The research presented in this paper is just the first step in the long period research plans. In the future work, we would like to improve the mechanisms used for control of the simulation process and the cell selection operation in order to reduce the number of sequential operations in the system as much as possible. We plan to apply KVDB-based data system, such as Apache HBase to improve the organization of agent data tables. Additionally, the generic environmental model should be improved to illustrate both individuals and environment activities, which allow us to provide our simulations in fully dynamic environments. Further experiments with larger crowd sizes and highly complex environments should be performed to improve the stability and efficiency of the developed simulation platform.

ACKNOWLEDGMENT

The inspiration for the research presented in this paper is the result of work in the IC1406 COST Action Horizon 2020 project cHiPSet "High-Performance Modelling and Simulation for Big Data Applications".

REFERENCES

- [1] T. J. Shields, K. E. Boyce and N. McConnell, "The behaviour and evacuation experiences of WTC 9/11 evacuees with self-designated mobility impairments," *Fire Safety Journal*, vol. 44, pp. 881-893, 2009.
- [2] P. F. Johnson, C. E. Johnson and C. Sutherland, "Stay or Go? Human Behavior and Decision Making in Bushfires and Other Emergencies," *Fire Technology*, vol. 48, pp. 137-153, 2012.
- [3] C. M. Henein and T. White, "Macroscopic effects of microscopic forces between agents in crowd models," *Physica A-Statistical Mechanics And Its Applications*, vol. 373, pp. 694-712, 2007.
- [4] D. Lee, J. H. Park and H. Kim, "A study on experiment of human behavior for evacuation simulation," *Ocean Engineering*, vol. 31, pp. 931-941, 2004.
- [5] R. L. Hughes, "The flow of human crowds", *Annual Review of Fluid Mechanics*, vol. 35, pp. 169-182, 2003.
- [6] L. Y. Jiao and Q. Y. Jiang, "Study on Emergency Management in Large-Social Activities Based on Behavior Modification Theory", *Progress in Safety Science and Technology, Vol. 8, Pts A And B*, Part a, pp. 451-454, 2010.
- [7] J. Was and K. Kulakowski, "Agent-Based Approach in Evacuation Modeling," *Agent and Multi-Agent Systems: Technologies and Application*, pp. 325-330, 2010.
- [8] Y. Q. Lin, I. Fedchenia, B. LaBarre, and R. Tomastik, "Agent-Based Simulation of Evacuation: An Office Building Case Study," *Pedestrian And Evacuation Dynamics 2008*, pp. 347-357, 2008.
- [9] M. H. Zaharia, F. Leon, C. Pal, and G. Pagu, "Agent-Based Simulation of Crowd Evacuation Behavior," in *Proceedings Of The 11th Wseas International Conference on Automatic Control, Modelling And Simulation*, pp. 529-533, 2009.
- [10] S. Sharma, H. Singh and A. Prakash, "Multi-Agent Modeling and Simulation of Human Behavior in Aircraft Evacuations," *IEEE Transactions On Aerospace And Electronic Systems*, vol. 44, pp. 1477-1488, 2008.
- [11] N. Zarboutis and N. Marmaras, "Design of formative evacuation plans using agent-based simulation," *Safety Science*, vol. 45, pp. 920-940, 2007.
- [12] T. Mao, H. Jiang, J. Li, Y. Zhang, S. Xia, and Z. Wang, "Parallelizing continuum crowds," in *Proceedings of the 17th ACM*

Symposium on Virtual Reality Software and Technology (ACM VRST) 2010, 231-234, 2010.

[13] G. Vigueras, M. Lozano, C. Perez, and J. M. Orduna, "A scalable architecture for crowd simulation: Implementing a parallel action server," *International Conference on Parallel Processing*, pp. 430-437, 2008.

[14] D. Chen, L. Wang, C. Bian, and X. Zhang, "A grid infrastructure for hybrid simulations," *Computer Systems Science and Engineering*, vol. 26, pp. 197-206, 2011.

[15] D. Chen, L. Wang, X. Wu, J. Chen, S. U. Khan, J. Kolodziej, M. Tian, F. Huang, and W. Liu, "Hybrid modelling and simulation of huge crowd over a hierarchical Grid architecture," *Future Generation Computer Systems*, vol. 29, pp. 1309-1317, 2013.

[16] Y. Wang, M. Lees and W. Cai, "Grid-based partitioning for large-scale distributed agent-based crowd simulation," in *Proceedings of the 2012 Winter Simulation Conference*, Berlin, Germany, pp. 2727-2738, 2012.

[17] E. Yilmaz, V. Isler and Y. Y. Cetin, "The virtual marathon: Parallel computing supports crowd simulations," *IEEE Computer Graphics and Applications*, vol. 29, pp. 26-33, 2009-01-01 2009.

[18] I. G. Georgoudas, P. Kyriakos, G. C. Sirakoulis, and I. T. Andreadis, "An FPGA implemented cellular automaton crowd evacuation model inspired by the electrostatic-induced potential fields," *Microprocessors and Microsystems*, vol. 34, pp. 285-300, 2010-01-01 2010.

[19] H. C. Yang, Z. B. Wang and J. S. Peng, "Production simulation using a distributed node-aware system," *IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 2057-2061, 2010.

[20] G. Pratz and L. Xing, "Monte Carlo simulation of photon migration in a cloud computing environment with MapReduce," *Journal of Biomedical Optics*, vol. 16, 2011-01-01 2011.

[21] A. Ciobanu and F. Ipate, "P system testing with parallel simulators - A survey," *Scalable Computing*, vol. 14, pp. 169-179, 2013-01-01 2013.

[22] J. Niu, S. Bai, E. Khosravi, and S. Park, "A Hadoop approach to advanced sampling algorithms in molecular dynamics simulation on cloud computing," *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 452-455, 2013.

[23] O. Seckic, C. Dorn and S. Dustdar, "Simulation-based modeling and evaluation of incentive schemes in crowdsourcing environments," *Confederated International Conferences on On the Move to Meaningful Internet Systems*, pp. 167-184, 2013.

[24] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Communications Of The Acm*, vol. 51, pp. 107-113, 2008.

[25] R. Laemmel, "Google's MapReduce programming model - Revisited," *Science Of Computer Programming*, vol. 70, pp. 1-30, 2008.

[26] J. Dean and S. Ghemawat, "MapReduce: A Flexible Data Processing Tool," *Communications Of The Acm*, vol. 53, pp. 72-77, 2010.

[27] M. Dou, J. Chen, D. Chen, X. Chen, Z. Deng, X. Zhang, K. Xu, and J. Wang, "Modeling and simulation for natural disaster contingency planning driven by high-resolution remote sensing images," *Future Generation Computer Systems*, 2013-01-01 2013.

[28] D. Grzonka, "The Analysis of OpenStack Cloud Computing Platform: Features and Performance", *Journal of Telecommunications and Information Technology*, 3/2015, pp. 52-57.

[29] Taylor, R. C Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics", *BMC Bioinformatics*, 2010 Supplement 12, Vol. 11, p1, 2010.

[30] S. Wolfram, "Theory and applications of cellular automata," *Advanced Series on Complex Systems, Singapore: World Scientific Publication*, 1986.

[31] S. Wolfram, *Cellular automata and complexity: collected papers* vol. 1: ISBN 0201626640, Addison-Wesley Reading, 1994.

[32] R. Guo, H. Huang and S. C. Wong, "A potential field approach to the modeling of route choice in pedestrian evacuation," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, p. P02010, 2013-01-01 2013.

[33] A. Byrski, M. Kisiel-Dorohinicki, "Agent-based model and computing environment facilitating the development of distributed computational intelligence systems", *Computational Science-ICCS 2009*, 2009.

[34] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems", *Human Factors Journal*, 37, 32-64, 1995



ANDRZEJ WILCZYŃSKI graduated with distinction in Computer Science from Cracow University of Technology, Poland, in 2015. He also studied at Fontys University of Applied Sciences, Netherlands, in 2012. Currently, he is a research and teaching assistant at Cracow University of Technology and a Ph.D. student at AGH University of Science and Technology, he works as a programmer as well. For more information, please visit: www.awilczynski.me. His e-mail address is: and.wilczynski@gmail.com.



JOANNA KOŁODZIEJ has graduated in Mathematics from the Jagiellonian University in Cracow in 1992, where she also obtained the PhD in Computer Science in 2004. She is employed at Cracow University of Technology as a professor. 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. Prof. Kołodziej is a Managing Editor of IJSSC Journal and serves as a EB member and guest editor of several peer-reviewed international journals. For more information, please visit: www.joannakołodziej.org.

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

-

Special Session

SIMULATION AND SELECTION OF EFFICIENT DECISION RULES IN BANK'S MANUAL UNDERWRITING PROCESS

Mikhail Konovalov
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia,
Email: mkonovalov@ipiran.ru

Rostislav Razumchik
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia,
Peoples' Friendship University of Russia,
Moscow, Russia
Email: rrazumchik@ipiran.ru

KEYWORDS

bank's underwriting, optimal rule, scheduling, dispatching, simulation

ABSTRACT

Bank's manual underwriting involves a group of underwriting inspectors, which perform a known set of procedures with the loan applications submitted by the borrowers, in order to determine the risk of providing a loan and eventually approve or disapprove it. Due to the fact that the evaluation process of applications must satisfy quality of service requirements usually set at legislator level and the bank resources are limited, one has to define such dispatching rules, that specify which application must be sent to which inspector and when in such a way that the requirements are met. This paper presents a case study of the application of "computer-aided scheduling" to the new problem of optimal management of applications, which is seen in the bank manual underwriting process. Here it is shown that the problem of optimal distribution of applications between the inspectors in the bank's manual underwriting can be represented as an optimal dispatching problem, commonly encountered in the distributed processing environment. We build the simulation model of the corresponding dispatching system and find best decision rule with the help of computational simulations. The realization of best decision-making is done by finding in a given set of dispatching rules the best one (either static or adaptive) for a given criterion and by estimating its parameters (if needed). By virtue of numerical examples it is shown how the quality of service requirements are met using different dispatching rules.

INTRODUCTION

The problem which is being considered in this paper frequently arises in the bank manual underwriting process and concerns algorithms, which are implemented in the bank information support systems and are used to manage incoming loan applications¹. In banking, underwriting is the process of approving or denying of a loan,

¹Documents that provide the essential financial and other kind of information about the borrower on which the bank bases the decision to lend.

based on the financial information and credit history of the borrower. An appropriate information support system is usually used for the the management of the underwriting process. Up to now there are two types of underwriting systems: automatic and manual. In an automatic underwriting system (credit scoring system), the information from the application is entered by a bank worker into a computer program which determines whether the borrower financially fits the loan conditions. Usually it takes from 5 minutes up to 5 hours to make a decision. If the loan is approved through an automatic system, the bank will proceed with it. If it is not, then the bank either rejects the application or hands it further for the manual underwriting.

In the manual underwriting² a bank worker (underwriting inspector) performs a set of procedures instead of a computer program and eventually determines the risk of providing a loan. In cases of high risks he can propose new conditions under which the loan can be provided. These procedures are specific for each bank and can be very different from the procedures run inside the automatic system. For example, they can include sending official requests to the federal state services, checking of the persistent arrears problems, evaluation of financial information, contacting internal bank services (such as security service). It is being reported that in case of manual underwriting it takes from 1 to 10 days to make a decision but this period heavily depends on the bank, application type and size of a loan. Despite the fact that in this case underwriting is done almost manually, a bank needs a supporting information system which will manage applications and track their status. Another goal of the system is to distribute applications among the available underwriting inspectors in order to keep them equally or unequally loaded³. But due to the fact that applications must be evaluated within the given time limits (deadlines), which are usually specified at a legislator level, the supporting information system must be able to make such dispatching decisions (i.e. specify which application must be sent to which inspector and when), that al-

²In general the manual underwriting is used for commercial (or business) loans.

³The objectives may be very different and usually depend on the bank's goals.

low the bank to meet the required deadlines. From the given rough description of the manual underwriting process, it is clear that the evaluation process of an application requires a random amount time even if the bank has well-established internal underwriting procedures. This fact in combination with the unpredictability of the application submission times and the requirement to meet the target deadlines makes the dispatching decisions complicated. The situation becomes even rougher as soon as one tries to take into consideration more and more details of the manual underwriting such as skills of the underwriting inspectors, working schedule, priorities of the applications, load balancing etc. It can also turn out that the deadlines and other quality of service (QoS) requirements, that may be specified at a legislator level (or within the bank) in order to increase its competitiveness, cannot be met by the bank. In the latter case the reason may not be in the bank's internal underwriting procedures, but in the deficit of the number of inspectors or in the lack of experienced staff. The true reason is not that easy to figure out because, for example, the given number of employees cannot cope with the flow of applications because the incoming applications are distributed among them in a wrong way, which can be optimized. It can be seen in practice, that when the bank realizes that it cannot cope with the current flow of applications, the dispatching decisions are taken under manual control of the executive manager, who decides when and who must evaluate which application. Eventually this leads to the accumulation of expired applications and inevitable consequences like penalties. Now most of the banks have an appropriate information support system for the manual underwriting and the bank's main goal is to tune it in a proper way so that the QoS requirements are met. In order to tune the system one has to answer the question: given that the bank understands its internal manual underwriting procedures and has relevant statistical information on how well (quantitatively and qualitatively) these procedures were performed for a certain period of time in the past, what is the best strategy (algorithm) to distribute the incoming applications between underwriting inspectors, which allows one to meet the QoS requirements? Despite a great generality of the question, from our point of view, the answer to it can be given by using computer-aided scheduling techniques and by seeing an analogy between the described information support system and the task/resource allocation problems typically met in the field of the distributed computations. The idea is that firstly one represents the problem of distribution of incoming applications among inspectors under given constraints (inspector skills, application priorities etc.) as a resource allocation problem (or a dispatching problem). Secondly one identifies QoS requirements that need to be met (for example, not more than 10% of overdue applications). Thirdly, one builds the simulation model of the system, which (at least to some extent) can reproduce the values of the chosen performance characteristics provided by the information support system currently oper-

ated by the bank. Fourthly, one finds the best possible solution by simulation and generates the improved dispatching rule. Finally, the bank, having implemented the new rule in the information support system, keeps tracking of the QoS requirements. As soon as it detects severe violations, the simulation model is re-built (if there were any changes in the underwriting procedures or staff) and run again with the updated historical data eventually generating the new rule, that is implemented in place of the previous one and used until the next violation.

Even though the idea is simple⁴ it has a number of drawbacks. The first step looks to be the most difficult one because of the high level of ambiguity: there are many ways in which manual underwriting procedures can be formalized. Some procedures can be left out of consideration, others cannot and it is hard to determine what the right granularity level is. Another aspect is the performance evaluation of the underwriting procedures. Inter alia this includes the understanding of performance characteristics of underwriting inspectors (how fast one copes with different stages of application evaluation process), estimation of the true times needed to fulfil the internal procedures and external official requests. The latter is not possible without enough historical data. Moreover the estimations of time frames can be done only on the probability basis. Finally, making the simulation results conform to the results achieved in real-life is another challenge which can be done only through trials and errors. Additionally, the need to re-build model each time the underwriting processes are changed, requires such a simulation framework in which models are built in an algorithmic manner. Given that simulation experiments are cheap compared to real-life implementation, trials and errors may have the price that can be paid. Our experience shows that in close cooperation with the bank these difficulties can be overcome at some abstraction level. Even if the abstraction level is high, which means that only basic procedures are being formalized, simulation can be advantageous, because the optimal decision rules may depend exactly on these basic procedures⁵. By building even a rough though consistent simulation model one obtains a basis to judge if the QoS requirements can be met without any changes in the underwriting procedures or staff number.

The case study presented in this paper demonstrates that the proposed idea indeed can be fulfilled. To our knowledge there are no studies on this topic in the literature. Here we demonstrate that the management (including assignment) of applications by inspectors during the manual underwriting process can be seen as a service process in a dispatching system. We give a short description of our algorithmic simulation framework and present the results of the numerical experiments based on syn-

⁴And not new. It is reported to be quite common approach in production scheduling in the field of industrial manufacturing. See, for example, Harmonosky and Robohn (1987).

⁵Of course, this is not always the case. But in the situation when one needs a more or less argumentative improved decision rule, this looks to be a feasible approach.

thetic data, which show how the QoS requirements of the manual underwriting process are met using different dispatching rules. Even though the comparison of the simulation results with the results from real-life experiments is not given here, one can see that in considered environment complex rules may be advantageous to simple ones. However the great increase in the rule's complexity does not lead to the prospective increase in the performance. Thus the appropriate rule is a matter of trade-off.

MANUAL UNDERWRITING PROCESS AS A DISPATCHING PROCESS

On a certain abstraction level one can see that the manual underwriting processes flow within one system, composed of typical objects: applications and inspectors. Typically there are several types of applications, that the bank works with and a pool of inspectors with different qualifications, which indicate the types of applications the inspector is allowed to evaluate. The manual underwriting process itself is just a sequence of actions that an inspector performs with each application. The number and sequence of actions depend on the type of the application and are usually fixed and are rarely changed within the bank. Due to the fact that the evaluation of the application requires contacting internal and external services, inspector is not busy with the evaluation of a single application during all his working time. It can be seen in practice that one inspector is evaluating several applications in parallel. The number of applications that an inspector can evaluate at the same time depends on his qualification level and rules of the bank. Each inspector is busy with the evaluation of the assigned applications strictly according to his working schedule: thus there are periods of unavailability when the applications are postponed until an inspector becomes available. The bank launches the underwriting process for each new application. Thus each application has a start time (when the application is submitted) and a finish time (when the application is approved or disapproved). In order to meet the quality of service requirements the bank has to appropriately control the objects of the system. But there is not too much in the system that can be controlled. For example, the bank cannot control the submission times of new applications; they arrive in stochastic manner. Time frames during which applications are evaluated are not deterministic and can vary significantly due to different reasons such as sudden illness of the inspector or unexpected delay at external service. Most of the unknown values can be estimated only on the probabilistic basis and the bank possesses only one major control option: set the rule according to which arriving applications are distributed between the inspectors.

The analogy between the described system and the dispatching system is apparent. In a dispatching system, each server (underwriting inspector) has its own queue and the task is to assign the arriving jobs (applications) to servers either immediately upon arrival or later, in or-

der to meet the objectives. Each job consists of several tasks (actions which inspector perform on the application), which have to be served in a prescribed sequence. The next subsection contains the detailed description of the dispatching system, which models key aspects of a manual underwriting process.

DESCRIPTION OF THE DISPATCHING SYSTEM

The dispatching system consists of N of servers without a dedicated queue each. There are M types of job flows that arrive at the system. Flows are independent and times between successive arrivals in each flow are i.i.d. random variables. A job within each flow has a deadline, consists of a number of tasks, which have to be served sequentially, one by one and a job is considered to be completed when all tasks it is comprised of are completed. Each task within a job is served in two steps. The first step is the preparation phase⁶, which does not require the processing time of the server. The second step is the service phase, when the server processes the task. Preparation and service times are considered to be i.i.d. random variables with cumulative distribution functions (CDF) D^I and D^{II} , respectively, which are considered to be known⁷. Thus each task \mathbb{T} can be described by a pair (D^I, D^{II}) , which defines how long a task is prepared and then served. Thus in order to introduce different task types one only needs to change CDF D^I and/or D^{II} . Jobs within one flow are homogeneous: a job \mathbb{J} in a flow can be described by a set $(d, \mathbb{T}_1, \dots, \mathbb{T}_k)$, where d is a deadline of a job, k is the total number of tasks within a job and \mathbb{T}_i is the task type⁸. Finally, each flow \mathbb{F} can be defined by a pair (D, \mathbb{J}) , where D is the CDF of the inter-arrival times of job \mathbb{J} . The joint arrival flow in the system can be described by a set $\mathbb{F} = (\mathbb{F}_1, \dots, \mathbb{F}_M)$.

Each server can handle a certain number of jobs in parallel and this number is called the server's capacity, which we denote by c . Servers' capacities can be different and are assumed to be known⁹. The server is considered to be busy when it has at least one job in service. Due to the fact that jobs consist of several tasks and each task is performed in two stages we have to specify how the service process goes on. When a job arrives at the server, its first task immediately starts getting prepared. This (preparation) time does not require server's processing power and though the server is considered to be busy it is in fact idle and ready to process other, already prepared tasks (if any are present). When one has finished

⁶The preparation phase is introduced in order to model times when the application is served by internal and external services. During these times the server (i.e. inspector) is not performing any work and is only considered to be busy. Thus it can process other tasks which are already prepared.

⁷We assume that these times can be estimated using historical data.

⁸Here one can observe clear resemblance with practice. If one considers the flow of a certain type of applications, when it is clear that applications are homogeneous, require the same number and type of actions from the inspector.

⁹This corresponds to the fact that an inspector can handle several applications at a time.

the preparation of the task and the server is idle, it starts processing the task and the service time is determined by server's speed r (which can be different for different servers) and task's service time v , and is equal to v/r . In the meanwhile the next task of the job starts getting prepared (which again does not require server's processing power). The server can process only one task at a time and if at the moment when the server becomes idle there are prepared tasks it starts to process a next task¹⁰. The speeds of the servers are assumed to be known. We also assume that server's availability is periodic¹¹ and denote by t_0 and t_1 the availability and unavailability periods correspondingly. Note that the preparation of each task does not depend on the server's availability (except for the fact that it can be started only when the server is available). But the server can become unavailable, even if it currently serves the task. The pre-emption of tasks is not allowed. Thus the type of a server can be described by the set $\mathbb{R} = (c, r, t_0, t_1)$ and the specification of the system's resources is given by $\mathbb{R} = (\mathbb{R}_1, \dots, \mathbb{R}_N)$. We again note that the set of parameters (\mathbb{F}, \mathbb{R}) is considered to be known a priori.

CONTROL IN THE SYSTEM AND THE COST FUNCTION

The dynamics of the described dispatching model depends on the dispatching rule for the incoming jobs¹². As each job is admitted into the system and each server can handle several jobs at a time, then the dispatching rule must specify how the server is assigned for the newly arriving job and how the server (upon service completion) chooses the next task to be served. Here a variety of options exists. For example, the assignment of the job to a server can be based on server's current utilization¹³ and the choice of the next task can depend on the total number of prepared tasks and on jobs' deadlines. In order to limit the number of options we assume that the system has a two-level hierarchical structure. All servers are grouped into K disjoint sets (clusters). Number of servers in the i -th cluster is n_i , $i = 1, \dots, K$, and servers within a cluster may be non-homogeneous¹⁴. Clearly, $\sum_{i=1}^K n_i = N$. These K clusters form the bottom level. The upper level of the system consists of a single entity called dispatcher. Each newly arriving job firstly goes to the dispatcher, which routes it immediately to one of the clusters. When a job arrives at the cluster and there are available servers in the cluster for this job, it is immediately assigned to

one of them. Otherwise it is kept in a virtual queue of the cluster until one of the appropriate servers becomes available. Note that once the job has entered a cluster it cannot leave it.

For such a two-level hierarchical structure one can introduce an agent based structure of a dispatching rule. Dispatcher agent A_0 makes decisions whereto route newly arriving jobs. Decisions are made at once. Agent A_i , $i = 1, \dots, K$, are responsible for assigning jobs within i -th cluster. Specifically the agent A_i consists of two agents $(A'_i; A''_i)$. Agent A'_i decides whereto route the job when it arrives at the i -th cluster. Due to the fact that at that moment all appropriate for the job servers may be busy, the agent may decide to put a job in a queue. Agent A''_i is responsible for assigning jobs, which are waiting in a queue, to servers each time when a server in the i -th cluster becomes available. Because a server may be busy with multiple jobs at a time, then upon service completion of a task there may be several other tasks ready for service. Agents A_{ki} are responsible for choosing the next task when server k in the i -th cluster finishes service of the previous task.

Thus a dispatching rule A consists of $(K+N+1)$ agents and symbolically can be written as

$$A = (A_0; A'; A''; \tilde{A}),$$

where $A' = (A'_1, \dots, A'_K)$, $A'' = (A''_1, \dots, A''_K)$. $\tilde{A} = (A_{ki}, i = 1, \dots, n_K; k = 1, \dots, K)$.

For the considered dispatching problem it is hardly possible to find analytically an optimal dispatching rule A . Following the common practice, we will use heuristic rules in conjunction with simulation in order to find the best rule in the given set of rules. Here we will present the results for the following common heuristics¹⁵: uniform random, least loaded first, first-in-first-out. Each dispatching rule A is assumed to be constructed from one or several of these heuristics¹⁶. In the next section we show the efficiency of different dispatching rules with respect to the cost function which is equal to the mean number of on-time jobs (i.e. mean number of jobs that were served within their deadlines). It is important to note that this cost function is not fair. Indeed it implies that some jobs may not be served at all (in case of high load) or may starve (i.e. may be served with severe deadline violations), which does not influence the average value of the cost function. A number of ways exist to make it fair (for example, one can introduce progressive penalties which grow with the growth of deadline violation times) but we don't consider them here.

OBTAINING THE BEST POSSIBLE SOLUTION

The optimization is based on the statistical simulation techniques. The problem considered in this paper lies in

¹⁵A review of dispatching policies up to 2011 can be found, for example, in Semchedine et al. (2011).

¹⁶For example, the agent A_0 may route newly arriving jobs either to a random cluster, or least loaded one.

¹⁰It means, that we consider only work-conserving disciplines.

¹¹This corresponds to working schedules of the inspectors.

¹²This directly corresponds to our assumption that the bank can control the underwriting process only by assigning newly submitting applications to inspectors.

¹³But in the considered problem it is unclear how to uniquely determine server's utilization. This is due to the fact that tasks which comprise a job are served in two stages: preparation phase and service phase. During the preparation phase the server either may process another task (if any) or may be free (though is still considered to be busy) if there are no other already prepared tasks.

¹⁴That is of different capacities and speeds.

the area of distributed computing and evolutionary computation. Not need to mention that in this area there are numerous research papers devoted to solving dynamic optimization problems using simulation which resulted in a variety of methods. Among the latest ones which include reviews on the topic one can refer, for example, to Kolodziej (2012); Doroudi et al. (2014); Broberg et al. (2006); Harchol-Balter et al. (1999). Below we describe in short our approach and highlight its peculiar features (for the details one can refer to Konovalov (2007)).

The solution of the optimization problem is found by building the simulation model and using adaptive optimization algorithms on simulated trajectories. For the first step we use our flexible simulation framework for job allocation problems in distributed processing systems. It allows one to build logical processes governing jobs' handling, is algorithmic and, to some extent, allows assembly of complex models from simple ones. Its formal description is based on the concept of communicating sequential processes introduced by C.A.R. Hoare (see Konovalov and Razumchik (2014); Konovalov (2014, 2007)).

The simulation model allows one to obtain the value of the cost function for any dispatching rule A . By carrying out series of experiments with different dispatching rules one can find the one which leads to the best value of the cost function. But there are reasons which may make this exhaustive search inapplicable. One reason is the time needed to obtain the stable value of the cost function. This time depends on the rate of convergence in the strong law of large numbers and may take too long. Exhaustive search becomes also difficult whenever the number of dispatching rules under test is infinite. We try to overcome these kind of difficulties by using adaptive search algorithms for partially observable Markov decision processes. Such algorithms use a single trajectory for each dispatching rule and tune its parameters in such a way that the probability of better values of the cost function is increased. For the detailed description of the algorithms used one can refer to Sragovich (2007); Konovalov (2007).

NUMERICAL EXAMPLE

Consider a bank which has 5 different business lines and thus the bank's manual underwriting process must handle 5 different types of loan applications. The specification of each application flow is given in Table 3. One can see that each flow is considered to be Poisson and on average the bank has 26 applications per day. We assume that the bank identifies 5 different actions (i.e. task types) that can be performed by an inspector when evaluating an application. The assumptions concerning the task types, including preparation and processing times, are stated in Table 1. We assume that each application type (i.e. job type) consists of 3 tasks and jobs differ from one another only by types of tasks they consist of (see Table 2). A total of 55 underwriting inspectors work in the bank and

have different qualifications¹⁷. Bank ranges inspectors by granting each one a certain category. We assume that there are 5 different categories (from 1 to 5), with # 1 denoting the highest one. In practice the category indicates which types of applications are available to the inspector. Usually the higher category of the inspector, the more important applications is allowed to evaluate. Within one category all inspectors are assumed to behave identically. The specification of each of the inspector's category is given in Table 4. One can see that, for example, inspectors belonging to the 1st category are able to handle 4 applications at a time, have 720 working minutes (12 hours) which are succeeded by 720 off-work minutes.

We also introduce an assumption that all inspectors are grouped into clusters (or teams). As mentioned in the previous section this artificial hierarchical view is introduced purely in order to make a decision process a little simpler. Table 5 shows the specification of each cluster.

Table 1: Specification of the task types

Task type	Time to prepare, D^I $D^I \sim \text{Pareto}(x_{min}, a)$	Service time, D^{II} $D^{II} \sim N(m, \sigma)$
\mathbb{T}_1	$x_{min} = 800, a = 6$ mean=960	$m = 960, \sigma = 20$
\mathbb{T}_2	$x_{min} = 480, a = 3$ mean=720	$m = 480, \sigma = 20$
\mathbb{T}_3	$x_{min} = 320, a = 3$ mean=480	$m = 240, \sigma = 10$
\mathbb{T}_4	$x_{min} = 160, a = 3$ mean=240	$m = 480, \sigma = 1$
\mathbb{T}_5	$x_{min} = 80, a = 3$ mean=120	$m = 120, \sigma = 1$

Table 2: Specification of the job (application) types

Job type	Types of task	deadline, d in min.
J_1	$\mathbb{T}_1, \mathbb{T}_2, \mathbb{T}_3$	8640
J_2	$\mathbb{T}_1, \mathbb{T}_2, \mathbb{T}_5$	7200
J_3	$\mathbb{T}_1, \mathbb{T}_3, \mathbb{T}_4$	5760
J_4	$\mathbb{T}_2, \mathbb{T}_3, \mathbb{T}_4$	5760
J_5	$\mathbb{T}_2, \mathbb{T}_3, \mathbb{T}_5$	4320

The cost function under consideration is the mean number of jobs served within their deadlines. We define several dispatching rules (see Table 6) and look for the one, which allows us to make the value of the cost function as high as possible. In Fig. 1 and Fig. 2. one can see the values of the cost function and corresponding values of the mean sojourn time (i.e. mean evaluation time for an arbitrary application) for each of the dispatching rules.

One can see from the figures that the best result one managed to achieve is little more than 86% of jobs served on time (dispatching rule A_5). But the dispatching rule A_1 , which is much easier to implement, is worse only by less than 1%. Moreover the rule A_4 , which does not take

¹⁷For example, security classification, experience with stock market clients, municipalities.

Table 3: Specification of the flow (application) types. Times between successive arrivals for each flow are exponential

Flow (application) type	Rate	Job (application) type	Suited servers (inspectors)
F_1	0,006 ≈ 9 per day	J_1	R_2, R_4
F_2	0,002 ≈ 3 per day	J_2	$R_1,$
F_3	0,003 ≈ 5 per day	J_3	R_1, R_3
F_4	0,001 ≈ 2 per day	J_4	R_4
F_5	0,005 ≈ 7 per day	J_5	R_2, R_5

Table 4: Specification of the servers (inspectors)

Server (inspector) type	Capacity, c	Speed, r	Schedule (on/off)	Total number
R_1	4	3	720/720	5
R_2	3	2.5	600/840	10
R_3	3	2	600/840	10
R_4	2	1.5	600/840	10
R_5	2	1	480/960	20

Table 5: Specification of the clusters (teams)

	Number of servers (inspectors) of type				
	R_1	R_2	R_3	R_4	R_5
Cluster 1	1	2	2	2	4
Cluster 2	1	3	4	0	0
Cluster 3	1	3	2	3	4
Cluster 4	1	1	2	3	2
Cluster 5	1	1	0	2	10

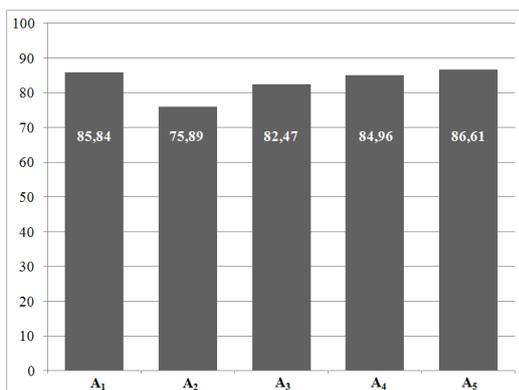


Figure 1: Percent of the jobs served within a deadline for each dispatching rule

into account type of incoming job, is worse by less than 2% than the best rule A_5 . Finally the simplest rule with random choice is worse by only 10% (approx. 10 hours). Here one can see that the complication of the dispatching rule does not lead to significant increase in the value of the cost function but may require too much for implementation (see, for example, Konovalov and Razumchik

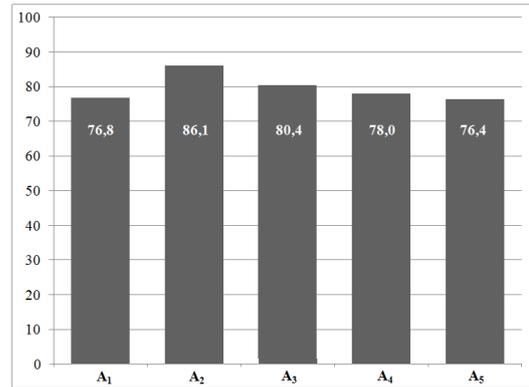


Figure 2: Mean evaluation time of an application (in hours) for each dispatching rule

(2014)). A simple heuristic rule (for example, decision based only on the number of applications) may lead to very good performance. As our experiments indicate the gain of the dispatching rule is very sensitive to the structure of the system (i.e. number of inspectors, number and types of applications etc.) and sometimes a change of a dispatching rule may lead to 20-30% performance increase.

SUMMARY

Here one has demonstrated that the bank's manual underwriting process can be played back in a large-scale dispatching system, which is a common mathematical model for various distributed processing systems. The goal of such modelling is to select a decision rule under which the bank's manual underwriting process will perform at a given level of quality of service. The idea, which was utilized to select an efficient decision rule, is well-known: perform search in a given set of heuristic rules using simulation. Due to the complexity of the system it is analytically intractable and this idea is apparently the only one which allows one to obtain an adequate solution. The special simulation framework and algorithms that we have used¹⁸ allowed us to carry simulation experiments with fairly large-scale systems (500 servers/inspectors) on a stand-alone PC. It is worth noticing that even a simple model of the manual underwriting process may be of an advantage for a bank because it allows to select an appropriate decision rules based on something more than intuition and expert's opinion. Generally speaking it is possible to build an adequate model of the bank's manual underwriting process, which will take into account not only those basic properties mentioned in this paper, but also many specific ones: skills of the inspectors, restricted access to applications, occasional events (illnesses), load balancing. Such model will not require much more computational resources and thus will not incur huge extra cost from the bank. Though in this paper the control is restricted only to the most

¹⁸Reader can refer to Konovalov and Razumchik (2014) and Konovalov (2007) for more details.

Table 6: Specification of the decision rules used

Decision rule	Rule for A_0	Rule for A'_k	Rule for A''_k	Rule for A_{ki}
A_1	minimum load per flow type ^a	minimum load per flow	FIFO	FIFO
A_2	random	random	random	random
A_3	minimum number of jobs	minimum number of jobs	FIFO	FIFO
A_4	minimum mean backlog ^b	minimum mean backlog	FIFO	FIFO
A_5	minimum load per flow type	minimum mean backlog with thresholds ^c	FIFO	FIFO

^aThis value is calculated as the mean total service time of all jobs in a cluster of the same type as the incoming one.

^bThis value is calculated as the mean total service time of all jobs in a cluster.

^cThis value is calculated as the product (mean total service time of all jobs in a cluster of the same type as the incoming one) × (threshold value). This is the analogue of the threshold policy used, for example, in Hyytia (2013).

common case – routing of applications, – one can also speculate whether the performance of the manual underwriting process depends on the staff structure (i.e. the number and type of inspectors and teams). Our experiments show that if one has control over this component as well, the performance of the underwriting processes can be improved even more.

Acknowledgements This work was supported by the Russian Foundation for Basic Research (grant 15-07-03406).

REFERENCES

- Haruhiko, S., Hiroaki, Sa. 2013. Online Scheduling in Manufacturing. A Cumulative Delay Approach. Springer-Verlag London.
- Min Hee Kim, Yeong-Dae Kim. 1994. Simulation-based real-time scheduling in a flexible manufacturing system. *Journal of Manufacturing Systems*. Vol. 13. Issue 2. Pp.85–93.
- Harmonosky, C.M., Robohn, S. F. 1991. Real-time scheduling in computer integrated manufacturing: a review of recent research. *International Journal of Computer Integrated Manufacturing*. Vol. 4. No. 6. Pp. 331–340.
- Konovalov, M., Razumchik, R. 2014. Simulation Of Task Distribution In Parallel Processing Systems. *Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems*. Pp. 657–663.
- Konovalov, M. G. 2014. Building a simulation model for solving scheduling problems of computing resources. *Systems and Means of Informatics*. Vol. 24. No. 4. Pp. 45–62. (in Russian)
- Konovalov, M. G. 2007. *Methods of Adaptive Information Processing and Their Applications*. Moscow: IPI RAN. (in Russian)
- Sragovich V.G. 2005. *Mathematical Theory Of Adaptive Control*. Singapore: World Scientific.
- Hyytia, E. 2013. Optimal Routing of Fixed Size Jobs to Two Parallel Servers. *INFOR: Information Systems and Operational Research*. Vol. 51. No. 4. Pp. 215–224.
- Konovalov M. G., Razumchik R. V. 2015. Approximate optimization of resource allocation strategy: the case of bank underwriting system. *Systems and Means of Informatics*. Vol. 25. No. 4. Pp. 31–51. (in Russian)

Semchedine, F., Bouallouche-Medjkoune, L., Aissani, D. 2011. Review: Task assignment policies in distributed server systems: A survey. *J. Netw. Comput. Appl.* Vol. 34. No. 4. Pp. 1123–1130.

Kolodziej, J. 2012. Evolutionary Hierarchical Multi-Criteria Metaheuristics for Scheduling in Large-Scale Grid Systems. *Studies in Computational Intelligence Series*. Vol. 419. Berlin-Heidelberg: Springer.

Doroudi, S., Hyytiö, E., Harchol-Balter, M. 2014. Value Driven Load Balancing. *Performance Evaluation*. Vol. 79. Pp. 306–327.

Broberg, J., Tari, Z., Zeephongsekul, P. 2006. Task assignment with work-conserving migration. *Journal of Parallel Computing*. Vol. 32. Pp. 808–830.

Harchol-Balter, M., Crovella, M., Murta, C. 1999. On Choosing a Task Assignment Policy for a Distributed Server System. *Journal of Parallel and Distributed Computing*. Vol. 59. Issue 2. Pp. 204–228.

AUTHOR BIOGRAPHIES

MIKHAIL KONOVALOV is a Doctor of Sciences in Technics and holds position of the principal scientist at Information Technologies Department at Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. His research activities are focused on adaptive control of random sequences, modelling and simulation of complex systems. His email address is mkonovalov@ipiran.ru.

ROSTISLAV RAZUMCHIK received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as a senior research fellow at Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS). Currently he holds the position of Head of the Information and Telecommunication System Modelling section at the FRC CSC RAS and associate professor position at Peoples' Friendship University of Russia. His current research activities are focused on queueing theory and its applications for performance evaluation of stochastic systems. His email address is rrazumchik@ipiran.ru

STATISTICAL CLASSIFICATION IN MONITORING SYSTEMS

Alexander A. Grusho, Nick A. Grusho and Elena E. Timonina
Institute of Informatics Problems,
Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences
Vavilova 44-2,
119333, Moscow, Russia
Email: grusho@yandex.ru

KEYWORDS

statistical decision functions, bans of probability measures, secure architectures, intelligent monitoring

ABSTRACT

The paper is devoted to the statistical classification problems. Repeated classification in control and monitoring systems is complicated by nonzero mistakes of traditional statistical decisions. At repeated applications of rules of statistical classification small probabilities of mistakes generate a large number of wrong decisions. At construction of monitoring systems of information security in computer systems wrong decisions are especially dangerous. Therefore for construction of secure architecture of control and monitoring systems it is necessary to look for nonconventional statistical decisions.

In finite set of words of finite length the ban is the word having zero probability of appearance. If the statistical criterion has the critical set consisting of only bans of supposed probability measure, the probability of wrong rejection of this measure is equal to zero. Therefore repeated application of such criterion won't generate to false alarms in monitoring systems.

In the paper we consider a case of statistical classification when classes are defined by finite sets of probability distributions on a space of infinite sequences. We use bans to define decision functions and prove conditions when these decisions produce no mistakes.

INTRODUCTION

The paper is devoted to the statistical classification problems. Statistical classification of data is often used in different mathematical modeling problems. However repeated classification in control and monitoring systems is complicated by nonzero mistakes of traditional statistical decisions. At repeated applications of rules of statistical classification small probabilities of mistakes generate a large number of wrong decisions (Axelson, 1999).

At construction of monitoring systems of information security in computer systems the wrong decisions are especially dangerous. Monitoring and control systems are widely explored in different directions (Socolov et al., 2013). One of the most important model of monitoring systems is a stochastic one. Therefore for construction

of secure architecture (Grusho et al., 2015a) of control and monitoring systems it is necessary to look for nonconventional statistical decisions. That is why we use a concept of a ban of probability measure in discrete probability space (Grusho and Timonina, 2011; Grusho et al., 2010).

In finite set of words of finite length the ban is the word having zero probability of appearance. If the statistical criterion has the critical set consisting of only bans of supposed probability measure, the probability of wrong rejection of this measure is equal to zero. Therefore repeated application of such criterion won't generate to false alarms in monitoring systems (Grusho et al., 2013a).

There are certain special applications of statistical methods defined by bans. For example in (Denisov, 2015) the search of inserted functional relations in random sequences is investigated. The important problem is to determine bans of considered measures. This problem was solved with help of statistical simulation of the analyzed measures (Grusho et al., 2013b). Such simulation helps to get consistent estimation of the set of bans.

At research of such criteria it is proved that under certain conditions there is consistency meaning that power function tends to 1 on each alternative (Grusho et al., 2013a). Conditions when power function becomes equal to 1 for all alternatives on a finite step are found (Grusho et al., 2014, 2015b).

In this paper we present the generalization on a case of statistical classification when classes are defined by finite sets of probability distributions on a space of infinite sequences.

The paper is structured as follows. Section 2 introduces definitions and previous results. In Section 3 the main results are proved. In Conclusion we shortly analyze future problems of construction of decision functions.

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 Tychonoff 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 be a projection of P on the first n coordinates of sequences from X^∞ . It is clear that for every $B_n \subseteq X^n$

$$P_n(B_n) = P(B_n \times X^\infty). \quad (1)$$

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

$$\Delta(P) = \lim_{n \rightarrow \infty} \Delta_n = \bigcap_{n=1}^{\infty} \Delta_n. \quad (2)$$

The set $\Delta(P)$ 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 (Grusho et al., 2014) of 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 (Grusho et al., 2014).

If \bar{x}_k is a ban of measure P_n then for every $k \leq s \leq n$ and for every \bar{x}_s sequence starting with \bar{x}_k we have

$$P_s(\bar{x}_s) = 0.$$

If there exists $\bar{x}_n \in X^n$ such that $P_n(\bar{x}_n) = 0$ then there exists the smallest ban. That is why further we say simply a ban of measure P .

Let on (X^∞, \mathcal{A}) probability measures $P^{(1)}, \dots, P^{(s)}$ be defined. As before we also define $P_n^{(i)}, D_n(P^{(i)}), \Delta_n(P^{(i)}), \Delta(P^{(i)}), i = 1, 2, \dots, s$.

Further under $\Lambda_n^{(i)} i = 1, 2, \dots, s$, we will understand a set of the smallest bans of measure $P_n^{(i)}$, which have lengths equal to n .

Let's construct a graph G on vertices $P^{(1)}, \dots, P^{(s)}$. Vertices $P^{(i)}$ and $P^{(j)}$ are connected by edge in graph G if and only if $\Delta(P^{(i)}) \cap \Delta(P^{(j)}) \neq \emptyset$. Let $Q^{(1)}, \dots, Q^{(r)}$ be sets of vertices in components of graph G , and for $i = 1, \dots, r, V_i$ be a set of indexes of vertices, including in $Q^{(i)}$. Denote

$$\Delta(V_i) = \bigcup_{j \in V_i} \Delta(P^{(j)}), i = \overline{1, r}.$$

For $x \in X^\infty$ denote $x|_n$ be a vector which includes the first n coordinates of the sequence x .

Then let's consider a sequence of decision functions $d_n(x|_n) = i, i = \overline{1, r}, n = 1, 2, \dots$

The basic problem considered in the paper is to find conditions when there exists such N that for all $n \geq N$ we can determine such decision function $d_n(x|_n)$ defined by bans that

$$P_n^{(i)}(d_n(x|_n) = j) = 1, \quad (3)$$

where $i = 1, 2, \dots, s, j = 1, \dots, r, i \in V_j$.

MATHEMATICAL RESULTS

Let $P^{(i)}, Q^{(i)}$ and V_i be defined as in Section 2. The solution for the basic problem is described in the next theorem.

Theorem 1. *There exists a sequence of $d_n(x|_n), n = 1, 2, \dots$, defined by bans, for which exists such N , that for every $n \geq N$ equations (3) are fulfilled if and only if*

$$\Delta(V_i) \cap \Delta(V_j) = \emptyset,$$

for $i \neq j, i, j = 1, \dots, r$.

The proof of the theorem 1 is based on several lemmas.

Let \bar{x}_k be the smallest ban of measure $P^{(i)}$. Then define $I_i(\bar{x}_k)$ be the elementary cylindrical set in X^∞ , which is generated by the vector \bar{x}_k .

Lemma 1. For every sequence $x \in I_i(\bar{x}_k)$ it follows that $x \notin \Delta(P^{(i)})$.

Proof. Suppose that there exists $x \in I_i(\bar{x}_k)$ that belongs to $\Delta(P^{(i)})$. From formula (2) it follows that $x \in \Delta_n(P^{(i)})$ for every $n = 1, 2, \dots$. By the definition of $\Delta_k(P^{(i)})$ the vector $x|_k$ defined by the first k coordinates of x belongs to the set $D_k(P^{(i)})$. Then

$$P_k(x|_k) > 0.$$

Besides

$$x|_k = \bar{x}_k,$$

that contradicts to supposition. The lemma 1 is proved.

Let's define the open set S_i :

$$S_i = \bigcup_{k=1}^{\infty} \bigcup_{\bar{x}_k \in \Lambda_k} I_i(\bar{x}_k). \quad (4)$$

From lemma 1 it follows that

$$S_i \cap \Delta(P^{(i)}) = \emptyset.$$

Lemma 2. The set S_i can be represented in the next form

$$S_i = X^\infty \setminus \Delta(P^{(i)}).$$

Proof. From

$$S_i \cap \Delta(P^{(i)}) = \emptyset$$

it follows that

$$S_i \subseteq X^\infty \setminus \Delta(P^{(i)}).$$

Let's assume that

$$x \in X^\infty \setminus \Delta(P^{(i)}).$$

If $x \in X^\infty \setminus \Delta(P^{(i)})$ then

$$x \notin \Delta(P^{(i)}) = \bigcap_{n=1}^{\infty} \Delta_n(P^{(i)}).$$

The sequence of sets $\{\Delta_n(P^{(i)})\}$ is not increasing. Then there exists n such that for every $t \geq n$ we have $x \notin \Delta_t(P^{(i)})$. That means that $P_t(x|_t) = 0$. Thus there exists the smallest ban \bar{x}_k such that $x \in I_i(\bar{x}_k)$, so $x \in S_i$. Lemma is proved.

Lemma 3.

$$\Delta(V_j) \cap \Delta(P^{(i)}) = \emptyset$$

if and only if

$$\Delta(V_j) \subseteq S_i.$$

Proof. From the condition of lemma 3 it follows that

$$\Delta(V_j) \subseteq X^\infty \setminus \Delta(P^{(i)}).$$

Then from lemma 2 $\Delta(V_j) \subseteq S_i$.

On the other hand if $\Delta(V_j) \subseteq S_i$, then

$$\Delta(V_j) \subseteq X^\infty \setminus \Delta(P^{(i)}),$$

and it follows that

$$\Delta(V_j) \cap \Delta(P^{(i)}) = \emptyset.$$

Lemma is proved.

Lemma 4. If

$$\Delta(V_j) \cap \Delta(P^{(i)}) = \emptyset$$

then $\exists N_i$ such that

$$\Delta(V_j) \subseteq \bigcup_{k=1}^{N_i} \bigcup_{\bar{x}_k \in \Lambda_k} I_i(\bar{x}_k).$$

Proof. Tychonoff product X^∞ is a compact space (Bourbaki, 1968) and therefore from an every infinite cover of a compact by open sets it is possible to select a finite cover. The closed set $\Delta(V_i)$ is a compact and $\Delta(V_j) \subseteq S_i$. That's why due to definition (4) there exists N such that

$$\Delta(V_j) \subseteq \bigcup_{k=1}^{N_i} \bigcup_{\bar{x}_k \in \Lambda_k} I_i(\bar{x}_k) = \sigma_{N_i}(i). \quad (5)$$

Lemma 4 is proved.

The set $\sigma_{N_i}(i)$ is a cylindrical set. Therefore it can be represented in the next form

$$\sigma_{N_i}(i) = C_{N_i}^{(i)} \times X^\infty,$$

where

$$C_{N_i}^{(i)} \subseteq X^{N_i}.$$

Lemma 5. For every $j = 1, \dots, r$,

$$\Delta(V_j) = \bigcap_{n=1}^{\infty} \bigcup_{i \in V_j} \Delta_n(P^{(i)}). \quad (6)$$

Proof. For every $i = 1, \dots, s$

$$\Delta(P^{(i)}) = \bigcap_{n=1}^{\infty} \Delta_n(P^{(i)}).$$

Union in formula (6) for every j is finite and sequence $\Delta_n(P^{(i)})$ for every i is non increasing. Then

$$\bigcup_{i \in V_j} \bigcap_{n=1}^{\infty} \Delta_n(P^{(i)}) = \bigcap_{n=1}^{\infty} \bigcup_{i \in V_j} \Delta_n(P^{(i)}).$$

The lemma 5 is proved.

Lemma 6. If $\forall t \neq j$,

$$\Delta(V_t) \cap \Delta(V_j) = \emptyset,$$

then $\exists N$:

$$\Delta_N(V_t) \cap \Delta_N(V_j) = \emptyset.$$

Proof. According to lemma 5

$$\bigcap_{n=1}^{\infty} \bigcup_{i \in V_t} \Delta_n(P^{(i)}) \cap \bigcap_{n=1}^{\infty} \bigcup_{i \in V_j} \Delta_n(P^{(i)}) = \emptyset,$$

and

$$\bigcap_{n=1}^{\infty} \{[\bigcup_{i \in V_t} \Delta_n(P^{(i)})] \cap [\bigcup_{i \in V_j} \Delta_n(P^{(i)})]\} = \emptyset.$$

Due to compactness of Tychonoff product (Bourbaki, 1968) it follows that $\exists N_{t,j}$ that for $\forall N \geq N_{t,j}$

$$\bigcap_{n=1}^N \bigcup_{i \in V_t} \Delta_n(P^{(i)}) \cap \bigcap_{n=1}^N \bigcup_{i \in V_j} \Delta_n(P^{(i)}) = \emptyset.$$

From monotonicity to n of

$$\bigcup_{i \in V_t} \Delta_n(P^{(i)})$$

and

$$\bigcup_{i \in V_j} \Delta_n(P^{(i)})$$

it follows that

$$\bigcup_{i \in V_t} \Delta_N(P^{(i)}) \cap \bigcup_{i \in V_j} \Delta_N(P^{(i)}) = \emptyset.$$

Lemma 6 is proved.

Corollary of Lemma 6. If true probability distribution of x is in $Q^{(i)}$ then from lemma 4 it follows that supports of all measures $P_N^{(j)}$, $j \notin V_i$, are covered by bans of true measure. That is every vector from $D_N(P^{(j)})$ includes bans of true measure.

Let's now prove the theorem 1. Let's define the decision function $d_N(\bar{x}_N)$ satisfying to conditions of the theorem 1. Note that defined earlier sets $C_{N^{(i)}}^{(i)}$, $i = 1, 2, \dots, s$, are determined by bans. Then they are defined by the smallest bans. Let's denote

$$N = \max_{i=1,2,\dots,s} N^{(i)}.$$

Consider the next matrix $M = \|\Lambda_j^{(i)}\|$ of size $s \times N$, where element of i -th row and j -th column is the set of smallest bans of the measure $P^{(i)}$ and of the length j .

For the observed random sequence x let's determine vectors $x|_1, \dots, x|_N$. For every one of these vectors compare its value with elements of certain column of matrix M . If any coincidence is found then let's mark that row where the vector is an element of the set of the smallest bans. It is clear that the true probability distribution cannot produce any coincidence. It follows from the fact that in x there are no bans of the true distribution.

That's why the row for the true distribution in M is not marked.

For every possible x and the row of M corresponding to the probability distribution different from the true distribution and located in other component of G with probability 1 we find a ban in x which belongs to that row. It follows from the corollary of lemma 6. That's why all rows of M corresponding to the probability distributions in different components from the true distribution are marked. Then we can define the decision function $d_N(x|_N)$ takes value j if in a component $Q^{(j)}$ there is a row i of matrix M which is unmarked. From lemma 6 it follows that this function is determined correctly for every x . Note that this function is defined by bans. The sufficiency is proved.

Let's prove the necessity. Let there exists N and a function $d_N(\bar{x}_N)$ defined by bans such that for all $i = 1, 2, \dots, s, j = 1, \dots, r,$

$$P_N^{(i)}(d_N(\bar{x}_N) = j) = 1,$$

when $i \in V_j.$

Define the set

$$B_N^{(j)} = \{\bar{x}_N : d_N(\bar{x}_N) = j\}$$

and denote

$$R_N^{(j)} = \bigcup_{i=1, i \neq j}^r \{\bar{x}_N : d_N(\bar{x}_N) = i\}.$$

Then for $P^{(i)} \in Q^{(j)}$

$$D_N(P^{(i)}) \subseteq B_N^{(j)}.$$

Due to the definition for all $i = 1, 2, \dots, s,$ we have

$$P_N^{(i)}(R_N^{(j)}) = 0.$$

Then due to definition

$$R_N^{(j)} \cap D_N(P^{(i)}) = \emptyset.$$

That's why

$$R_N^{(j)} \times X^\infty \subseteq \sigma(i).$$

As

$$P_N^{(l)}(B_N^{(t)}) = 1$$

for $t \neq j$ and $l \in V_t,$ $d(\bar{x}_N)$ is correctly defined, then

$$D_N(P^{(l)}) \cap D_N(P^{(i)}) = \emptyset$$

and

$$D_N(P^{(l)}) \subseteq R_N^{(j)}.$$

Then for $t \neq j$

$$\Delta_N(V_j) \cap \Delta_N(V_t) = \emptyset.$$

It follows that

$$\Delta(V_j) \cap \Delta(V_t) = \emptyset.$$

The theorem 1 is proved.

From the theorem 1 we can get the main result of (Grusho et al., 2016).

Theorem 2. *Let $P^{(1)}, \dots, P^{(s)}$ are such that for all $i, |Q^{(i)}| = 1,$ then there exists N and a function $d_N(\bar{x}_N)$ defined by bans such that for all $i = 1, 2, \dots, s,$*

$$P_N^{(i)}(d_N(\bar{x}_N) = i) = 1,$$

if and only if for all pairs $i, j, i \neq j, i = 1, 2, \dots, s, j = 1, 2, \dots, s,$

$$\Delta(P^{(i)}) \cap \Delta(P^{(j)}) = \emptyset.$$

Under the considered conditions we can define decision functions $g_n(x|_n),$ which are determined by the conditions

$$g_n(x|_n) = i,$$

where i is the $\min\{j\}$ such that

$$x|_n \in D_n(V_j).$$

It is possible to prove that there exists N when functions satisfy equation

$$P_N^{(i)}(g_N(x|_N) = j) = 1$$

for all $i \in V_j.$

But searching bans is equivalent to signature analysis, which proved to be quick.

It is interesting to compare these types of decision functions. Let's define rooted trees containing admissible trajectories of random data $x \in X^\infty$ produced under the probability distribution $P^{(i)}.$ The root of every tree means the first element of admissible sequence. Every infinite branch of the tree uniquely defines a sequence $x \in X^\infty$ in the same way at it is usually done in m -arc tree, edges show possible ways of development of the random sequence. All admissible parts of branches of the length n define the set $D_n(P^{(i)})$ and $X^n \setminus D_n(P^{(i)})$ is a set of all bans of the length $n.$ The set $D_n(P^{(i)})$ can be represented by trees of the height $n.$ All bans of the length n also can be represented by trees.

If we use the smallest bans for classification we use reduced vectors which have lengths less or equal $n.$ We can see smallest bans on the $D_n(P^{(i)})$ -trees. When we use the smallest bans then the number of steps to check that the vector $x|_n$ does not belong to $D_n(P^{(i)})$ demands less or equal then n steps. That's why we say about decision functions defined by bans.

No marks in the matrix M means that $x|_n$ may belong to $D_n(P^{(i)}).$ But all vectors $x|_n$ have to possess bans of all other measures when $n \geq N,$ because $x|_n$ doesn't belong to supports of measures in other components of the graph $G.$ If $n < N$ it is possible to make a mistake. It means that probability distribution in another component may be admitted as a part of the true one.

CONCLUSION

Conditions under which the statistical classification of random infinite sequence is reduced to consideration of finite space of vectors of finite length N are found in the article. The offered approach allows to generalize the

found conditions on a case when some classes of probability measures contain an infinite number of measures.

There is an open problem of constructive estimation of parameter N . If this problem was solved, from conditions on supports of measures in infinite spaces it would be possible to pass to conditions on supports of measures in finite spaces. In certain cases this problem is effectively solved. However the overall picture isn't visible yet.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grant 15-07-02053).

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. 2010. "Problems of modeling in the analysis of covert channels". *Lecture Notes in Computer Science* LNCS 6258. Heidelberg, Germany, 118–124.
- Grusho, A. and E. Timonina. 2011. "Prohibitions in discrete probabilistic statistical problems". *Discrete Mathematics and Applications* 21, No.3, 275–281.
- Grusho, A., N. Grusho and E. Timonina. 2013a. "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. 2013b. "Statistical Techniques of Bans Determination of Probability Measures in Discrete Spaces". *Informatics and Applications* 7, No.1, 54–57.
- Grusho, A., N. Grusho and E. Timonina. 2014. "Generation of Probability Measures with the Given specification of the Smallest Bans". In *Proceedings of 28th European Conference on Modelling and Simulation* (May 27-30, 2014, Brescia, Italy). Digitaldruck Pirrot GmbH, Dudweiler, Germany, 565–569.
- Grusho, A., N. Grusho, E. Timonina and S. Shorgin. 2015a. "Possibilities of Secure Architecture Creation for Dynamically Changing Information Systems". *Systems and Means of Informatics* 25, No.3, 78–93.
- Grusho, A., N. Grusho and E. Timonina. 2015b. "Quality of tests defined by bans". In *Proceedings of the 16th Applied Stochastic Models and Data Analysis International Conference (ASMDA2015) with Demographics 2015 Workshop* (30 June 4 July 2015, Piraeus, Greece). Edt. Christos H Skiadas, 289–295.
- Grusho, A., N. Grusho and E. Timonina. 2016. "Properties of decision functions defined by bans". *Journal of Mathematical Sciences* (to appear).
- Denisov, O. 2015. "Statistical Methods of Search for Coordinate Set on which a Random Vector Has Bans". *Applied Discrete Mathematics* No.2, 5–20.
- Socolov, B., M. Okhtilev, S. Potryasaev and Yu. Merkurjev. 2013. "Multi-model Description of Monitoring and Control Systems of Natural and Technological Objects". *Information Technology and Management Science* 16, No.1, 11–17.
- Prokhorov, U.V., and U.A. Rozanov. 1993. *Theory of probabilities*. Science, Moscow.

AUTHOR BIOGRAPHIES

ALEXANDER A. GRUSHO, Professor (1993), Doctor of Science in physics and mathematics (1990). He is Head of laboratory in Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences and Professor of Moscow State University.

Research interests: probability theory and mathematical statistics, information security, discrete mathematics, computer sciences.

His email is grusho@yandex.ru.

NICK A. GRUSHO has graduated from the Moscow Technical University. He is Candidate of Science (PhD) in physics and mathematics. At present he works as senior scientist at Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

Research interests: probability theory and mathematical statistics, information security, simulation theory and practice, computer sciences.

His email is info@itake.ru.

ELENA E. TIMONINA has graduated from the Moscow Institute of Electronics and Mathematics and obtained the Candidate degree (PhD) in physics and mathematics (1974). She is Doctor in Technical Science (2005), Professor (2007). Now she works as leading scientist in Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

Research interests: probability theory and mathematical statistics, information security, cryptography, computer sciences.

Her email is eltimon@yandex.ru.

TWO-SIDED TRUNCATIONS OF INHOMOGENEOUS BIRTH-DEATH PROCESSES

Yacov Satin,
Anna Korotysheva, Ksenia Kiseleva,
Galina Shilova and Elena Fokicheva
Vologda State University
S.Orlova, 6, Vologda, Russia

Alexander Zeifman
Vologda State University,
Vologda, Russia
IPI FRC CSC RAS;
ISEDT RAS

Victor Korolev
Moscow State University,
Moscow, Russia
IPI FRC CSC RAS

KEYWORDS

Inhomogeneous birth-death processes; queueing models; two-sided uniform approximation bounds

ABSTRACT

We consider a class of inhomogeneous birth-death queueing models and obtain uniform approximation bounds of two-sided truncations. Some examples are considered. Our approach to truncations of the state space can be used in modeling information flows related to high-performance computing.

INTRODUCTION

It is well known that explicit expressions for the probability characteristics of stochastic birth-death queueing models can be found only in a few special cases. Therefore, the study of the rate of convergence as time $t \rightarrow \infty$ to the steady state of a process is one of two main problems for obtaining the limiting behavior of the process. If the model is Markovian and stationary in time, then, as a rule, the stationary limiting characteristics provide sufficient or almost sufficient information about the model. On the other hand, if one deals with inhomogeneous Markovian model then, in addition, the limiting probability characteristics of the process must be approximately calculated. The problem of existence and construction of limiting characteristics for time-inhomogeneous birth and death processes is important for queueing and some other applications, see for instance, [1], [3], [5], [8], [15], [16]. General approach and related bounds for the rate of convergence was considered in [13]. Calculation of the limiting characteristics for the process via truncations was firstly mentioned in [14] and was considered in details in [15], uniform in time bounds have been obtained in [17].

As a rule, the authors dealt with the so-called north-west truncations (see also [9]), namely they studied the truncated processes with the same first states

$0, 1, \dots, N$ In the present paper we consider a more general approach and deal with truncated processes on state space $N_1, N_1 + 1, \dots, N_2$ for some natural $N_1, N_2 > N_1$.

Let $X = X(t)$, $t \geq 0$ be a birth and death process (BDP) with birth and death rates $\lambda_n(t)$, $\mu_n(t)$ respectively.

Let $p_{ij}(s, t) = Pr \{X(t) = j | X(s) = i\}$ for $i, j \geq 0$, $0 \leq s \leq t$ be the transition probability functions of the process $X = X(t)$ and $p_i(t) = Pr \{X(t) = i\}$ be the state probabilities.

Throughout the paper we assume that

$$P(X(t+h) = j | X(t) = i) = \begin{cases} q_{ij}(t)h + \alpha_{ij}(t, h) & \text{if } j \neq i, \\ 1 - \sum_{k \neq i} q_{ik}(t)h + \alpha_i(t, h) & \text{if } j = i, \end{cases} \quad (1)$$

where all $\alpha_i(t, h)$ are $o(h)$ uniformly in i , i. e. $\sup_i |\alpha_i(t, h)| = o(h)$. Here all $q_{i,i+1}(t) = \lambda_i(t)$, $i \geq 0$, $q_{i,i-1}(t) = \mu_i(t)$ $i \geq 1$, and all other $q_{ij}(t) \equiv 0$.

The probabilistic dynamics of the process is represented by the forward Kolmogorov system of differential equations:

$$\begin{cases} \frac{dp_0}{dt} = -\lambda_0(t)p_0 + \mu_1(t)p_1, \\ \frac{dp_k}{dt} = \lambda_{k-1}(t)p_{k-1} - (\lambda_k(t) + \mu_k(t))p_k + \mu_{k+1}(t)p_{k+1}, \quad k \geq 1. \end{cases} \quad (2)$$

By $\mathbf{p}(t) = (p_0(t), p_1(t), \dots)^\top$, $t \geq 0$, we denote the column vector of state probabilities and by $A(t) = (a_{ij}(t))$, $t \geq 0$ the matrix related to (2). One can see that $A(t) = Q^\top(t)$, where $Q(t)$ is the intensity (or infinitesimal) matrix for $X(t)$.

We assume that all birth and death intensity functions $\lambda_i(t)$ and $\mu_i(t)$ are linear combinations of a finite number of functions which are locally integrable on $[0, \infty)$. Moreover, we suppose that

$$\lambda_n(t) \leq \Lambda_n \leq L < \infty, \quad \mu_n(t) \leq \Delta_n \leq L < \infty, \quad (3)$$

$$\mathbf{z}(t) = V(t, 0)\mathbf{z}(0) + \int_0^t V(t, \tau)\mathbf{f}(\tau) d\tau, \quad (8)$$

where $V(t, z)$ is the Cauchy operator of (5), see, for instance [13].

We have $\|\mathbf{f}(t)\|_{1D} = d_{i-1}\mu_i(t) + d_{i+1}\lambda_i(t) \leq d_{i-1}\Delta_i + d_{i+1}\Lambda_i$ for almost all $t \geq 0$. On the other hand, if we put

$$\beta_k(t) = \begin{cases} \lambda_k(t) + \mu_{k+1}(t) + \frac{d_{k+1}}{d_k}\lambda_{k+1}(t) + \frac{d_{k-1}}{d_k}\mu_k(t), & k < i-1 \\ \lambda_{i-1}(t) + \mu_i(t) + \frac{d_{i+1}}{d_{i-1}}\lambda_i(t) + \frac{d_{i-2}}{d_{i-1}}\mu_{i-1}(t), & k = i-1 \\ \lambda_i(t) + \mu_{i+1}(t) + \frac{d_{i+2}}{d_{i+1}}\lambda_{i+1}(t) + \frac{d_{i-1}}{d_{i+1}}\mu_i(t), & k = i \\ \lambda_k(t) + \mu_{k+1}(t) + \frac{d_{k+2}}{d_{k+1}}\lambda_{k+1}(t) + \frac{d_k}{d_{k+1}}\mu_k(t), & k > i. \end{cases} \quad (9)$$

then one has

$$\|B(t)\|_{1D} = \sup_{k \geq 0} \beta_k(t) \leq 4L - \alpha(t),$$

for almost all $t \geq 0$.

Then $\mathbf{f}(t)$ and $B(t)$ are bounded and locally integrable on $[0, \infty)$ as vector function and operator function in l_{1D} respectively.

Now we have the following bound for the logarithmic norm $\gamma(B(t))$ in l_{1D} :

$$\gamma(B)_{1D} = \gamma(DB(t)D^{-1})_1 = -\inf_{k \geq 0} (\alpha_k(t)) = -\alpha(t), \quad (10)$$

in accordance with (7), see detailed discussion in our previous papers [4], [5], [15], [17].

Hence

$$\|V(t, s)\|_{1D} \leq e^{-\int_s^t \alpha(\tau) d\tau}. \quad (11)$$

Suppose now that there exist positive M and α such that

$$e^{-\int_s^t \alpha(\tau) d\tau} \leq M e^{-\alpha(t-s)}, \quad (12)$$

for any $0 \leq s \leq t$. Then $X(t)$ is exponentially weakly ergodic in $1D$ norm.

Put now $\mathbf{z}(0) = 0$ (i. e., $\mathbf{p}(0) = \mathbf{e}_i$). Then we have

$$\begin{aligned} \|\mathbf{z}(t)\|_{1D} &\leq \|V(t, 0)\|_{1D} \|\mathbf{z}(0)\|_{1D} + \\ &+ \int_0^t \|V(t, s)\|_{1D} \|\mathbf{f}(s)\|_{1D} ds \leq \\ &\leq \int_0^t M e^{-\alpha(t-s)} \|\mathbf{f}(s)\|_{1D} ds \leq \\ &\leq \frac{1}{\alpha} M (d_{i-1}\Delta_i + d_{i+1}\Lambda_i). \end{aligned} \quad (13)$$

On the other hand

$$\|\mathbf{z}\|_{1D} = (d_0 + \dots + d_{i-1})p_0 + (d_1 + \dots + d_{i-1})p_1 + \dots + d_{i-1}p_{i-1} + d_{i+1}p_{i+1} + (d_{i+1} + d_{i+2})p_{i+2} + \dots$$

Denote

$$g_k = \sum_{j=k}^{i-1} d_j, \quad G_k = \sum_{j=i+1}^k d_j.$$

Then we have

$$\begin{aligned} \|\mathbf{z}(t)\|_{1D} &= \|D\mathbf{z}(t)\| = \sum_{k < i} p_k(t) g_k + \\ \sum_{k > i} p_k(t) G_k &\geq \begin{cases} p_k(t) g_k, & k < i \\ p_k(t) G_k, & k > i \end{cases} \end{aligned} \quad (14)$$

Hence $p_k(t) \leq \begin{cases} \frac{\|\mathbf{z}(t)\|_{1D}}{g_k}, & k < i \\ \frac{\|\mathbf{z}(t)\|_{1D}}{G_k}, & k > i \end{cases}$ and we obtain the following statement.

Theorem 1. Let a BDP $X(t)$ with rates $\lambda_k(t)$ and $\mu_k(t)$ be given. Assume that there exists a sequence $\{d_k\}$ such that (12) is fulfilled. Then $X(t)$ is exponentially weakly ergodic in $1D$ norm and the following bound holds

$$p_k(t) \leq \begin{cases} \frac{M(d_{i-1}\Delta_i + d_{i+1}\Lambda_i)}{\alpha g_k}, & k < i \\ \frac{M(d_{i-1}\Delta_i + d_{i+1}\Lambda_i)}{\alpha G_k}, & k > i \end{cases}, \quad (15)$$

for any k .

TWO-SIDED TRUNCATIONS

Consider truncated BDP on state space $N_1, N_1 + 1, \dots, N_2$ with intensities $\lambda_k^*(t) = \lambda_k(t)$ if $N_1 \leq k < N_2$, and $\mu_k^*(t) = \mu_k(t)$ if $N_1 < k \leq N_2$ and suppose other birth and death rates equal to zero. We will denote by $A^*(t)$, $\mathbf{p}^*(t)$ and so on the correspondent characteristics of truncated BDP.

We have for the truncated process the following equation

$$\frac{d\mathbf{p}^*}{dt} = A^*(t)\mathbf{p}^*(t), \quad (16)$$

instead of (4). Now, the property $\mathbf{p}^*(t) \in \Omega$ for any $t \geq 0$ allows to put $p_i^*(t) = 1 - \sum_{j \neq i} p_j^*(t)$, for arbitrary fixed i . Then we obtain from (16)

$$\frac{d\mathbf{z}^*}{dt} = B^*(t)\mathbf{z}^*(t) + \mathbf{f}^*(t). \quad (17)$$

instead of (5).

Rewrite (17) in the form:

$$\frac{d\mathbf{z}^*}{dt} = B(t)\mathbf{z}^*(t) + (B^*(t) - B(t))\mathbf{z}^*(t) + \mathbf{f}^*(t). \quad (18)$$

Then we have the following equality for the solutions of (5) and (18):

$$\begin{aligned} \mathbf{z}(t) - \mathbf{z}^*(t) &= V(t, 0)(\mathbf{z}(0) - \mathbf{z}^*(0)) + \\ &+ \int_0^t V(t, s)(B(s) - B^*(s))\mathbf{z}^*(s) ds + \\ &+ \int_0^t V(t, s)(\mathbf{f}(s) - \mathbf{f}^*(s)) ds. \end{aligned} \quad (19)$$

We will suppose that $\mathbf{z}(0) = \mathbf{z}^*(0) = 0$ (i. e., $\mathbf{p}(0) = \mathbf{p}^*(0) = \mathbf{e}_i$ or $X(0) = X^*(0) = i$), where $N_1 < i < N_2$. Then $\mathbf{f}(s) = \mathbf{f}^*(s)$, for any s .

Hence we have

$$\mathbf{z}(t) - \mathbf{z}^*(t) = \int_0^t V(t, s)(B(s) - B^*(s))\mathbf{z}^*(s) ds, \quad (20)$$

and

$$\begin{aligned} & ((B(s) - B^*(s)) \mathbf{z}^*(s)) = \\ & (0, \dots, 0, \mu_{N_1} p_{N_1}^*, -\mu_{N_1} p_{N_1}^*, 0, \dots, \\ & 0, -\lambda_{N_2} p_{N_2}^*, \lambda_{N_2} p_{N_2}^*, 0, \dots)^\top. \end{aligned} \quad (21)$$

Let $\{d_k^*\}$ be a sequence of positive numbers such that there exist positive M^* and α^* such that

$$e^{-\int_s^t \alpha^*(\tau) d\tau} \leq M^* e^{-\alpha^*(t-s)}, \quad (22)$$

for any $0 \leq s \leq t$, instead of (12), where

$$\alpha^*(t) = \min \alpha_k^*(t), \quad (23)$$

and

$$\alpha_k^*(t) = \begin{cases} \lambda_k^*(t) + \mu_{k+1}^*(t) - \frac{d_{k+1}}{d_k} \lambda_{k+1}^*(t) - \frac{d_{k-1}}{d_k} \mu_k^*(t), & k < i-1 \\ \lambda_{i-1}^*(t) + \mu_i^*(t) - \frac{d_{i+1}}{d_{i-1}} \lambda_i^*(t) - \frac{d_{i-2}}{d_{i-1}} \mu_{i-1}^*(t), & k = i-1 \\ \lambda_i^*(t) + \mu_{i+1}^*(t) - \frac{d_{i+2}}{d_{i+1}} \lambda_{i+1}^*(t) - \frac{d_{i-1}}{d_{i+1}} \mu_i^*(t), & k = i \\ \lambda_k^*(t) + \mu_{k+1}^*(t) - \frac{d_{k+2}}{d_{k+1}} \lambda_{k+1}^*(t) - \frac{d_k}{d_{k+1}} \mu_k^*(t), & k > i. \end{cases} \quad (24)$$

Hence we obtain

$$\begin{aligned} & \|(B(s) - B^*(s)) \mathbf{z}^*(s)\|_{1D} \leq \\ & |g_{N_1-1} + g_{N_1}| \mu_{N_1}(s) p_{N_1}^*(s) + \\ & |G_{N_2+1} + G_{N_2}| \lambda_{N_2}(s) p_{N_2}^*(s) \leq \\ & \leq 2g_{N_1-1} \Delta_{N_1} p_{N_1}^*(s) + 2G_{N_2+1} \Lambda_{N_2} p_{N_2}^*(s). \end{aligned} \quad (25)$$

$$\text{Put } g_k^* = \sum_{j=k}^{i-1} d_j^* \text{ and } G_k^* = \sum_{j=i+1}^k d_j^*.$$

Instead of (15) we have now

$$p_k^*(t) \leq \begin{cases} \frac{M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*)}{\alpha^* g_k^*}, & k < i \\ \frac{M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*)}{\alpha^* G_k^*}, & k > i \end{cases}. \quad (26)$$

Therefore (20), (25) and (26) imply the bound

$$\begin{aligned} & \|\mathbf{z}(t) - \mathbf{z}^*(t)\|_{1D} \leq \\ & \frac{2M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*)}{\alpha \alpha^*} \\ & \cdot \left(\frac{g_{N_1-1} \Delta_{N_1}}{g_{N_1}^*} + \frac{G_{N_2+1} \Lambda_{N_2}}{G_{N_2}^*} \right). \end{aligned} \quad (27)$$

Let

$$d = \min(d_{i-1}, d_{i+1}), \quad W = \inf \left(\frac{g_k}{k}, \frac{d}{i}, \frac{G_k}{k} \right). \quad (28)$$

We have the inequalities:

$$\begin{aligned} & |p_i - p_i^*| \leq |p_0 - p_0^*| + |p_{i-1} - p_{i-1}^*| + \\ & |p_{i+1} - p_{i+1}^*| + \dots \leq \frac{1}{d} \|\mathbf{z}(t) - \mathbf{z}^*(t)\|_{1D}, \end{aligned} \quad (29)$$

$$\|\mathbf{p}(t) - \mathbf{p}^*(t)\| \leq \frac{2}{d} \|\mathbf{z}(t) - \mathbf{z}^*(t)\|_{1D}, \quad (30)$$

$$\begin{aligned} & 2\|\mathbf{z}\|_{1D} \geq 1 \frac{d_1 + \dots + d_{i-1}}{1} p_1 + \dots + \\ & (i-1) \frac{d_{i-1}}{i-1} p_{i-1} + i \frac{d}{i} p_i + (i+1) \frac{d_{i+1}}{i+1} p_{i+1} + \\ & (i+2) \frac{d_{i+1} + d_{i+2}}{i+2} p_{i+2} + \dots \geq W \|\mathbf{p}\|_{1E}. \end{aligned} \quad (31)$$

Theorem 2. Let birth-death processes $X(t)$ and $X^*(t)$ be such that (12) and (22) hold. Let $\mathbf{p}(0) = \mathbf{p}^*(0) = e_i$ (i. e., $X(0) = X^*(0) = i$). Then the following bounds hold:

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\| \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*)}{d \alpha \alpha^*} \\ & \cdot \left(\frac{g_{N_1-1} \Delta_{N_1}}{g_{N_1}^*} + \frac{G_{N_2+1} \Lambda_{N_2}}{G_{N_2}^*} \right), \end{aligned} \quad (32)$$

and

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*)}{W \alpha \alpha^*} \\ & \cdot \left(\frac{g_{N_1-1} \Delta_{N_1}}{g_{N_1}^*} + \frac{G_{N_2+1} \Lambda_{N_2}}{G_{N_2}^*} \right). \end{aligned} \quad (33)$$

Corollary 1. Let under assumptions of Theorem 2 $N_2 = \infty$. Then the following bounds hold

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\| \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*) g_{N_1-1} \Delta_{N_1}}{d \alpha \alpha^* g_{N_1}^*}, \end{aligned} \quad (34)$$

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*) g_{N_1-1} \Delta_{N_1}}{W \alpha \alpha^* g_{N_1}^*}, \end{aligned} \quad (35)$$

for any $i > N_1$.

Corollary 2. Let under assumptions of Theorem 2 $N_1 = 0$. Then the following bounds hold

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\| \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*) G_{N_2+1} \Lambda_{N_2}}{d \alpha \alpha^* G_{N_2}^*}, \end{aligned} \quad (36)$$

$$\begin{aligned} & \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} \leq \\ & \frac{4M M^*(\Delta_i d_{i-1}^* + \Lambda_i d_{i+1}^*) G_{N_2+1} \Lambda_{N_2}}{W \alpha \alpha^* G_{N_2}^*}, \end{aligned} \quad (37)$$

for any $i < N_2$.

EXAMPLES

Example 1. Consider a birth-death process with periodical birth and death rates:

$$\begin{aligned}\lambda_k(t) &= \lambda(t) = 10 + \cos t \\ \mu_k(t) &= 1 + \sin t, 0 < k < 1000 \\ \mu_k(t) &= 24 + \sin t, k \geq 1000.\end{aligned}$$

Let $i = 1000$.
Firstly we put

$$\dots, d_{997} = 1.5^2, d_{998} = 1.5, d_{999} = 1, \\ d_{1001} = 1.5, d_{1002} = 1.5^2, d_{1003} = 1.5^3, \dots;$$

and

$$\dots, d_{997}^* = 4, d_{998}^* = 2, d_{999}^* = 1, \\ d_{1001}^* = 2, d_{1002}^* = 4, d_{1003}^* = 8, \dots$$

Then we obtain

$$\begin{aligned}M = 1, \alpha = \frac{1}{2}, M^* = 1, \alpha^* = \frac{1}{2}; \\ d = 1.5, W = \frac{1.5}{1001}; \\ \Lambda_i = 11, \Delta_i = 25.\end{aligned}$$

Hence Theorem 2 implies the following bounds:

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-12}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-9},\end{aligned}$$

for $N_1 = 350, N_2 = 1650$.

Let now $\{d_k^*\}$ be the same sequence, and let $\{d_k\}$ be the such that

$$\dots, d_{997} = 1.1^2, d_{998} = 1.1, d_{999} = 1, \\ d_{1001} = 1.1, d_{1002} = 1.1^2, d_{1003} = 1.1^3, \dots$$

Then one has

$$M = 1, \alpha = \frac{1}{2}, d = 1.1,$$

and the following bounds hold

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-14}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-11},\end{aligned}$$

for $N_1 = 800, N_2 = 1200$.

Example 2. Consider a birth-death process with periodical birth and death rates:

$$\begin{aligned}\lambda_k(t) &= \lambda(t) = 10 + \cos t \\ \mu_k(t) &= \mu(t) = 1 + \sin t, 0 < k < 10^6, \\ \mu_k(t) &= 24 + \sin t, k \geq 10^6.\end{aligned}$$

Put $i = 10^6$.

We consider the similar sequences $\{d_k\}$ and $\{d_k^*\}$.
Namely, let

$$\dots, d_{999997} = 1.5^2, d_{999998} = 1.5, d_{999999} = 1, \\ d_{1000001} = 1.5, d_{1000002} = 1.5^2, d_{1000003} = 1.5^3, \dots;$$

and

$$\dots, d_{999997}^* = 4, d_{999998}^* = 2, d_{999999}^* = 1, \\ d_{1000001}^* = 2, d_{1000002}^* = 4, d_{1000003}^* = 8, \dots$$

For the first sequence $\{d_k\}$ we have $d = 1.5$ and the following bounds hold:

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-12}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-6},\end{aligned}$$

if $N_1 = 99350, N_2 = 100650$.

Let $\{d_k\}$ be the such that

$$\dots, d_{999997} = 1.1^2, d_{999998} = 1.1, d_{999999} = 1, \\ d_{1000001} = 1.1, d_{1000002} = 1.1^2, d_{1000003} = 1.1^3, \dots$$

Then we have $d = 1.1$ and the following bound holds

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-14}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-8},\end{aligned}$$

for $N_1 = 90800, N_2 = 100200$.

Example 3. Consider a birth-death process with large periodical birth and death rates:

$$\begin{aligned}\lambda_k(t) &= \lambda(t) = 1000(10 + \cos t) \\ \mu_k(t) &= 1000(1 + \sin t), 0 < k < 10^3 \\ \mu_k(t) &= 1000(24 + \sin t), k \geq 10^3.\end{aligned}$$

Let $i = 10^3$.

Consider the same sequences $\{d_k\}$ and $\{d_k^*\}$. For the first sequence $\{d_k\}$ we have $d = 1.5$ and

$$\Lambda_i = 11000, \Delta_i = 25000, \Delta_{N_1} = 2000, \Lambda_{N_2} = 11000.$$

Hence the following bounds hold:

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-6}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-3},\end{aligned}$$

if $N_1 = 350, N_2 = 1650$.

For the second sequence $\{d_k\}$ we have $d = 1.1$ and the bounds:

$$\begin{aligned}\|\mathbf{p}(t) - \mathbf{p}^*(t)\| &\leq 10^{-8}, \\ \|\mathbf{p}(t) - \mathbf{p}^*(t)\|_{1E} &\leq 10^{-5},\end{aligned}$$

for $N_1 = 800, N_2 = 1200$.

CONCLUSIONS

In this paper we considered a class of inhomogeneous birth-death queueing models and obtained uniform approximation bounds of two-sided truncations. Such approximations can be used in studying the information flows related to high-performance computing. The development of methodology for other classes of inhomogeneous Markovian queueing models seems to be a promising direction of research.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research, projects no. 15-01-01698, 15-07-05316, 15-37-20851, and by Ministry of Education and Science.

REFERENCES

- [1] Beim, G. K., Hobbs, B. F. (1988). Analytical simulation of water system capacity reliability: 2. A Markov Chain Approach and verification of the models // *Water Resources Research*, **24**(9), 1445–1458.
- [2] Daleckij, Ju.L., Krein, M.G. (1974). Stability of solutions of differential equations in Banach space. *Amer. Math. Soc. Transl.* **43**.
- [3] Di Crescenzo A., Nobile, A. G. (1995). Diffusion approximation to a queueing system with time dependent arrival and service rates // *Queueing Syst.*, **19**, 41–62.
- [4] E. A. Van Doorn, A. I. Zeifman, T. L. Panfilova (2010). Bounds and asymptotics for the rate of convergence of birth-death processes // *Th. Prob. Appl.*, **54**, 97–113.
- [5] B. L. Granovsky, A. I. Zeifman (2004). Nonstationary Queues: Estimation of the Rate of Convergence // *Queueing Syst.* **46**, 363–388.
- [6] Knessl C. (2000). Exact and asymptotic solutions to a pde that arises in time-dependent queues // *Adv. Appl. Probab.*, **32**, 256–283.
- [7] Knessl C., Yang Y. P. (2002). An exact solution for an $M(t)/M(t)/1$ queue with time-dependent arrivals and service // *Adv. Appl. Probab.*, **40**, 233–248.
- [8] Mandelbaum A., Massey W. (1995). Strong approximations for time-dependent queues // *Math. Oper. Research*, **20**, 33–64.
- [9] Masuyama, H. (2015). Continuous-time block-monotone Markov chains and their block-augmented truncations // *arXiv preprint arXiv:1511.04669*.
- [10] P. R. Parthasarathy, B. Krishna Kumar (1991). Density-dependent birth and death processes with state-dependent immigration // *Mathematical and Computer Modelling*, **15**, 11–16.
- [11] Y. Satin, A. Zeifman, A. Korotysheva, K. Kiseleva, V. Korolev (2015). On Truncations For A Class Of Finite Markovian Queueing Models // *ECMS 2015 Proceedings* edited by: Valeri M. Mladenov, Petia Georgieva, Grisha Spasov, Galidiya Petrova European Council for Modeling and Simulation. doi:10.7148/2015-0626.
- [12] Tweedie R. L. (1998). Truncation approximations of invariant measures for Markov chains // *J. Appl. Probab.*, **35**, 517–536.
- [13] A. I. Zeifman (1995). Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes // *Stoch. Proc. Appl.*, **59**, 157–173.
- [14] Zeifman, A.I. (1988). Truncation error in a birth and death system // *USSR Computational Mathematics and Mathematical Physics*, **28**(6), 210–211.
- [15] A. Zeifman, S. Leorato, E. Orsingher, Ya. Satin, G. Shilova (2006). Some universal limits for nonhomogeneous birth and death processes // *Queueing Syst.*, **52**, 139–151.

[16] Zeifman A., Satin Y., Panfilova T. (2013). Limiting characteristics for finite birth-death-catastrophe processes // *Mathematical biosciences*, **245**, 96–102.

[17] A. Zeifman, Ya. Satin, V. Korolev, S. Shorgin (2014). On truncations for weakly ergodic inhomogeneous birth and death processes // *International Journal of Applied Mathematics and Computer Science*, **24**, 503–518.

[18] A. I. Zeifman, A. V. Korotysheva, V. Yu. Korolev, Ya. A. Satin (2016). Truncation bounds for approximations of inhomogeneous continuous-time Markov chains // *Th. Prob. Appl.*, submitted.

AUTHOR BIOGRAPHIES

YACOV SATIN is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. His email is yacovi@mail.ru.

ANNA KOROTYSHEVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. Her email is a_korotysheva@mail.ru.

KSENIA KISELEVA is PhD student, Vologda State University. Her email is ksushakiseleva@mail.ru.

GALINA SHILOVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Head of Department of Mathematics, Vologda State University. Her email is shgn@mail.ru.

ELENA FOKICHEVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. Her email is eafokicheva2007@yandex.ru.

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, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; principal scientist, Institute of Socio-Economic Development of Territories, Russian Academy of Sciences. His email is a_zeifman@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, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

ASYMPTOTIC EXPANSIONS FOR THE DISTRIBUTION FUNCTION OF THE SAMPLE MEDIAN CONSTRUCTED FROM A SAMPLE WITH RANDOM SIZE

Vladimir E. Bening, Victor Yu. Korolev
Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University;
IPI FRC CSC RAS

Alexander I. Zeifman
Vologda State University,
IPI FRC CSC RAS;
ISEDT RAS

KEYWORDS

Sample median; sample with random size; asymptotic expansion; Student distribution; Cauchy distribution; Laplace distribution.

ABSTRACT

Statistical regularities of the information flows in contemporary communication, computational and other information systems are characterized by the presence of the so-called “heavy tails”. The outlying observations make the traditional moment-type location estimators inaccurate. In this case the robust median-type location estimators are preferable. On the other hand, the random character of the intensity of the flow of informative events results in that the available sample size (traditionally this is the number of observations registered within a certain time interval) is random. The randomness of the sample size crucially changes the asymptotic properties of the estimators. In the paper, asymptotic expansions are obtained for the distribution function of the sample median constructed from a sample with random size. A general theorem on the asymptotic expansion is proved for this case. The cases of the Laplace, Student and Cauchy distributions are considered. Special attention is paid to the situations in which the heavy-tailed distributions (Cauchy, Laplace) are inherent in both the original sample and the asymptotic regularities of the sample median (Student, Laplace) due to the randomness of the sample size. This approach can be successfully used for big data mining and analysis of information flows in high-performance computing.

INTRODUCTION

Statistical regularities of the information flows in contemporary communication, computational and other information systems are characterized by the presence of the so-called “heavy tails”. The outlying observations make the traditional moment-type location estimators inaccurate. As is known, in this case the robust median-type estimators are preferable. On

the other hand, the random character of the intensity of the flow of informative events results in that the available sample size (traditionally this is the number of observations registered within a certain time interval) is random. The randomness of the sample size crucially changes the asymptotic properties of the estimators, see, e. g., [12], [3].

In the paper, asymptotic expansions (a. e.) are obtained for the distribution function (d. f.) of the sample median constructed from a sample with random size. The cases of the Laplace, Student and Cauchy distributions are considered. These results are further development of the research presented in [6], [1], [12], [15], [16], [7], [8], [4], [5].

Special attention is paid to the situations in which the heavy-tailed distributions (Cauchy, Laplace) are inherent in both the original sample and the asymptotic regularities of the sample median (Student, Laplace) due to the randomness of the sample size.

We use the following notation: \mathbb{R} and \mathbb{N} are the sets of real and natural numbers, respectively, $\Phi(x)$ and $\varphi(x)$ are the d. f. of the standard normal law and its density.

Let X_1, X_2, \dots, X_n be independent identically distributed random variables with the common d. f. $F(x - \theta)$ and probability density $p(x - \theta)$, where θ is the unknown location parameter to be estimated from the sample X_1, X_2, \dots, X_n . By $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ we denote the order statistics constructed from the original observations X_1, X_2, \dots, X_n . Let M_n be the sample median (see, e. g., [10], [18], [9]), that is,

$$M_n = \begin{cases} X_{(m+1)}, & n = 2m + 1, \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}), & n = 2m, m \in \mathbb{N}. \end{cases} \quad (1.1)$$

The first-order asymptotic properties of the sample median M_n are well known (see, e. g., the book [18], Theorem 5.3.2 on p. 313, or the book [17], p. 81). Namely, if $F(0) = \frac{1}{2}$ and $p(0) > 0$, then, as $n \rightarrow \infty$,

$$\sup_{x \in \mathbb{R}} |P_\theta(\sqrt{n}(M_n - \theta) < x) - \Phi(2xp(0))| \rightarrow 0, \quad (1.2)$$

$$E_\theta(M_n - \theta)^2 = (2p(0))^{-2} n^{-1} + o(n^{-1}). \quad (1.3)$$

The second-order asymptotic properties of the sample median were considered in [9]. Recall the main results of that paper. For this purpose, first, formulate the regularity conditions imposed in [9] on the density $p(x)$.

Condition 1.1. The density $p(x)$ is symmetric around zero, i. e., $p(-x) = p(x)$, $x \in \mathbb{R}$, and $p(0) > 0$.

Condition 1.2. The density $p(x)$ has three continuous bounded derivatives in some neighborhood of zero of the form $(0, \delta)$, $\delta > 0$.

Condition 1.3. There exist constants $C > 0$ and $\alpha > 0$ such that the d. f. $F(x)$ satisfies the inequality

$$1 - F(x) \leq Cx^{-\alpha}, \quad x > 0.$$

For example, note that these regularity conditions are satisfied by the Cauchy distribution with the density

$$p(x) = [\pi(1 + x^2)]^{-1}, \quad x \in \mathbb{R}, \quad (1.4)$$

and the Laplace distribution with the density

$$p(x) = \frac{1}{2}e^{-|x|}, \quad x \in \mathbb{R}. \quad (1.5)$$

For the Laplace distribution the sample median M_n coincides with the maximum likelihood estimator of the parameter θ (see, e. g., [9]).

In what follows we will use the notation

$$k = [n/2], \quad p_0 = p(0) > 0, \quad p_1 = p'(0+), \quad p_2 = p''(0+),$$

where $[\cdot]$ denotes the integer part of a number.

Theorem 1.1 [9]. 1. Let the density $p(x)$ satisfy the regularity conditions 1.1 and 1.2. Then

$$\begin{aligned} P_\theta(2p_0\sqrt{2k}(M_n - \theta) < x) &= \Phi(x) + \varphi(x)\frac{p_1x|x|\sqrt{2}}{8p_0\sqrt{k}} + \\ &+ \varphi(x)\frac{x}{8k}\left(3 + x^2 + \frac{x^2p_2}{6p_0^3} - \frac{x^4p_1^2}{8p_0^4}\right) + O(n^{-3/2}) \end{aligned}$$

uniformly in $x \in \mathbb{R}$

2. If the regularity conditions 1.1–1.3 hold, then

$$\begin{aligned} E_\theta(M_n - \theta)^2 &= \frac{1}{8p_0^2k} - \frac{p_1}{8p_0^4\sqrt{\pi}k^{3/2}} - \\ &- \frac{1}{16p_0^2k^2}\left(3 + \frac{p_2}{4p_0^3} - \frac{15p_1^2}{16p_0^4}\right) + O(n^{-5/2}). \end{aligned}$$

Corollary 1.1. 1. For the Laplace distribution (1.5) we have

$$\begin{aligned} P_\theta(\sqrt{2k}(M_n - \theta) < x) &= \Phi(x) - \varphi(x)\frac{x|x|\sqrt{2}}{4\sqrt{k}} + \\ &+ \varphi(x)\frac{x(18 + 10x^2 - 3x^4)}{48k} + O(n^{-3/2}), \\ E_\theta(M_n - \theta)^2 &= \frac{1}{2k} + \frac{1}{\sqrt{\pi}k^{3/2}} - \frac{1}{16k^2} + O(n^{-5/2}). \end{aligned}$$

2. For the Cauchy distribution (1.4) we have

$$\begin{aligned} P_\theta(2\sqrt{2k}(M_n - \theta) < \pi x) &= \Phi(x) + \\ &+ \varphi(x)\frac{x[9 + x^2(3 - \pi^3)]}{24k} + O(n^{-3/2}), \end{aligned}$$

$$E_\theta(M_n - \theta)^2 = \frac{\pi^2}{8k} + \frac{\pi^2(\pi^2 - 6)}{32k^2} + O(n^{-5/2}).$$

It is easy to see that

$$\begin{aligned} \frac{1}{\sqrt{k}} &= \frac{\sqrt{2}}{\sqrt{n}} + O(n^{-3/2}), \quad \frac{1}{k} = \frac{2}{\sqrt{n}} + \frac{1 + (-1)^{n+1}}{n^2} + O(n^{-3}), \\ \frac{1}{k^{3/2}} &= \frac{2^{3/2}}{n^{3/2}} + O(n^{-5/2}), \quad \frac{1}{k^2} = \frac{4}{n^2} + O(n^{-3}). \end{aligned}$$

Hence, the assertions of Theorem 1.1 and Corollary 1.1 can be rewritten as

$$\begin{aligned} P_\theta(2p_0\sqrt{2k}(M_n - \theta) < x) &= \Phi(x) + \varphi(x)\frac{p_1x|x|\sqrt{2}}{4p_0\sqrt{n}} + \\ &+ \varphi(x)\frac{x}{4n}\left(3 + x^2 + \frac{x^2p_2}{6p_0^3} - \frac{x^4p_1^2}{8p_0^4}\right) + O(n^{-3/2}), \\ E_\theta(M_n - \theta)^2 &= \frac{1}{4p_0^2n} - \frac{p_1\sqrt{2}}{4p_0^4\sqrt{\pi}n^{3/2}} + \\ &+ \frac{1}{4p_0^2n^2}\left(\frac{(-1)^{n+1} - 5}{2} - \frac{p_2}{4p_0^3} + \frac{15p_1^2}{16p_0^4}\right) + O(n^{-5/2}). \end{aligned}$$

For the Laplace distribution (1.5) we obtain the asymptotic relations

$$\begin{aligned} P_\theta(\sqrt{2k}(M_n - \theta) < x) &= \Phi(x) - \\ &- \varphi(x)\frac{x|x|}{2\sqrt{n}} + \varphi(x)\frac{x(18 + 10x^2 - 3x^4)}{24n} + O(n^{-3/2}), \\ E_\theta(M_n - \theta)^2 &= \frac{1}{n} + \frac{2\sqrt{2}}{\sqrt{\pi}n^{3/2}} + \frac{1}{2n^2}\left[\frac{1}{2} + (-1)^{n+1}\right] + O\left(\frac{1}{n^{5/2}}\right), \end{aligned}$$

whereas for the Cauchy distribution (1.4) we have

$$\begin{aligned} P_\theta(2\sqrt{2k}(M_n - \theta) < \pi x) &= \Phi(x) + \\ &+ \varphi(x)\frac{x[9 + x^2(3 - \pi^3)]}{12n} + O(n^{-3/2}), \\ E_\theta(M_n - \theta)^2 &= \frac{\pi^2}{4n} + \frac{\pi^2(\pi^2 + (-1)^{n+1} - 5)}{8n^2} + O(n^{-5/2}). \end{aligned}$$

ASYMPTOTIC EXPANSIONS FOR THE DISTRIBUTION FUNCTION OF THE SAMPLE MEDIAN CONSTRUCTED FROM A SAMPLE WITH RANDOM SIZE

In classical problems of mathematical statistics, the size of the available sample, i. e., the number of available observations, is traditionally assumed to be deterministic. In the asymptotic settings it plays the role of infinitely increasing known parameter. At the same time, in practice 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. Therefore, 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. For example, this is so in high-frequency financial statistics where the number of events in a limit order book during a time unit essentially depends on

the intensity of order flows [15]. In these cases the number of available observations as well as the observations themselves are unknown beforehand and should be treated as random to avoid underestimation of risks or error probabilities.

Therefore it is quite reasonable to study the asymptotic behavior of general statistics constructed from samples with random sizes for the purpose of construction of suitable and reasonable asymptotic approximations. As this is so, an appropriate *non-random* centering and normalization of the statistics under consideration must be used to obtain reasonable approximation to the distribution of the basic statistics. Otherwise the approximate distribution becomes random itself and, for example, the problem of evaluation of quantiles or significance levels becomes senseless.

In asymptotic settings, statistics constructed from samples with random sizes are special cases of random sequences with random indices. The randomness of indices usually leads to that the limit distributions for the corresponding random sequences are heavy-tailed even in the situations where the distributions of non-randomly indexed random sequences are asymptotically normal see, e. g., [2], [3], [13]. For example, if a statistic which is asymptotically normal in the traditional sense, is constructed on the basis of a sample with random size having negative binomial distribution, then instead of the expected normal law, the Student distribution with power-type decreasing heavy tails appears as an asymptotic law for this statistic.

As regards sample median constructed from a sample with random size, in [12] it was shown that, if the sample size has the geometric distribution, then, instead of the normal law expected in the classical situation (see Theorem 1.1), the actual asymptotic distribution of the sample median is the Student law with two degrees of freedom defined by the d. f.

$$S_2(x) = \frac{1}{2} \left(1 + \frac{x}{\sqrt{2+x^2}} \right), \quad x \in \mathbb{R}.$$

This distribution has so heavy tails that the moments of orders $\delta \geq 2$ do not exist.

Consider a problem setting that is traditional for mathematical statistics. Let random variables $N_1, N_2, \dots, X_1, X_2, \dots$, be defined on one and the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Assume that for each $n \geq 1$ the random variable N_n takes only natural values and is independent of the sequence X_1, X_2, \dots of independent identically distributed random variables. Let $T_n = T_n(X_1, \dots, X_n)$ be a statistic, that is, a measurable function of X_1, \dots, X_n . For every $n \geq 1$ define the random variable T_{N_n} as

$$T_{N_n}(\omega) = T_{N_n(\omega)}(X_1(\omega), \dots, X_{N_n(\omega)}(\omega))$$

for each $\omega \in \Omega$. The random variable T_{N_n} so defined is referred to as a statistic constructed from the sample with random size N_n . As this is so, $n \in \mathbb{N}$ is the “infinitely large” parameter required to make the asymptotic settings reasonable. For better understanding, n may be treated as the “mean” or “expected” or “most probable” value of N_n .

Now we formulate the condition that determines the a. e. for the statistic T_n under the non-random sample size n .

Condition 2.1. There exist constants $l \in \mathbb{N}$, $\mu \in \mathbb{R}$, $\sigma > 0$, $\alpha > l/2$, $\gamma \geq 0$, $C_1 > 0$, a differentiable d. f. $F(x)$ and bounded differentiable functions $f_j(x)$, $j = 1, \dots, l$ such that

$$\sup_x \left| \mathbb{P}(\sigma n^\gamma (T_n - \mu) < x) - F(x) - \sum_{j=1}^l \frac{f_j(x)}{n^{j/2}} \right| \leq \frac{C_1}{n^\alpha}, \quad n \in \mathbb{N}.$$

The following condition determines the a. e. for the d. f. of the normalized random sample size N_n .

Condition 2.2. There exist constants $m \in \mathbb{N}$, $\beta > m/2$, $C_2 > 0$, 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

$$\sup_{x \geq 0} \left| \mathbb{P}\left(\frac{N_n}{g(n)} < x\right) - H(x) - \sum_{i=1}^m \frac{h_i(x)}{n^{i/2}} \right| \leq \frac{C_2}{n^\beta}, \quad n \in \mathbb{N}.$$

In [7] the following statement was proved.

Theorem 2.1. Let the statistic $T_n = T_n(X_1, \dots, X_n)$ and the random sample size N_n satisfy Conditions 2.1 and 2.2, respectively. Then there exists a constant $C_3 > 0$ such that

$$\begin{aligned} \sup_x \left| \mathbb{P}(\sigma g^\gamma(n)(T_{N_n} - \mu) < x) - G_n(x) \right| &\leq \\ &\leq C_1 \mathbb{E} N_n^{-\alpha} + \frac{C_3 + C_2 D_n}{n^\beta}, \end{aligned}$$

where

$$D_n = \sup_x \int_{1/g(n)}^{\infty} \left| \frac{\partial}{\partial y} \left(F(xy^\gamma) + \sum_{j=1}^l \frac{f_j(xy^\gamma)}{(yg(n))^{j/2}} \right) \right| dy$$

and the a. e. $G_n(x)$ is defined by the formula $G_n(x) =$

$$\begin{aligned} &= \int_{1/g(n)}^{\infty} F(xy^\gamma) dH(y) + \sum_{i=1}^m \frac{1}{n^{i/2}} \int_{1/g(n)}^{\infty} F(xy^\gamma) dh_i(y) + \\ &\quad + \sum_{j=1}^l \frac{1}{g^{j/2}(n)} \int_{1/g(n)}^{\infty} \frac{f_j(xy^\gamma)}{y^{j/2}} dH(y) + \\ &\quad + \sum_{j=1}^l \sum_{i=1}^m \frac{1}{n^{i/2} g^{j/2}(n)} \int_{1/g(n)}^{\infty} \frac{f_j(xy^\gamma)}{y^{j/2}} dh_i(y). \end{aligned}$$

With the account of Theorem 1.1 it is easy to see that the sample median M_n satisfies Condition 2.1 with

$$\gamma = \frac{1}{2}, \quad \alpha = \frac{3}{2}, \quad l = 2, \quad \mu = \theta, \quad \sigma = 2p_0\sqrt{2}, \quad g(n) = \sqrt{k}, \quad (2.1)$$

$$F(x) = \Phi(x), \quad f_1(x) = \varphi(x) \frac{p_1 x |x| \sqrt{2}}{4p_0},$$

$$f_2(x) = \varphi(x) \frac{x}{4} \left(3 + x^2 + \frac{x^2 p_2}{6p_0^3} - \frac{x^4 p_1^2}{8p_0^4} \right). \quad (2.2)$$

In the same way as Lemma 5.1 was proved in [7], it can be shown that there exists a constant $D > 0$ such that

$$D_n \leq D, \quad n \in \mathbb{N}. \quad (2.3)$$

In [8] a similar theorem was obtained for a non-normalized statistic under the following regularity condition:

Condition 2.3. There exist constants $l \in \mathbb{N}$, $\alpha > l/2$, $C_1 > 0$, a differentiable d. f. $G(x)$ and bounded differentiable functions $g_i(x)$, $i = 1, \dots, l$ such that

$$\sup_x \left| \mathbb{P}(T_n < x) - G(x) - \sum_{i=1}^l \frac{g_i(x)}{n^{i/2}} \right| \leq \frac{C_1}{n^\alpha}, \quad n \in \mathbb{N}.$$

Theorem 2.2. Let Conditions 2 ([10]) and 2.3 hold. Then

$$\sup_x \left| \mathbb{P}(T_{N_n} < x) - G_n(x) \right| \leq C_1 \mathbb{E}N_n^{-\alpha} + \frac{2C_2}{n^\beta} \sup_x \sum_{i=1}^l |g_i(x)|,$$

where the function $G_n(x)$ has the form $G_n(x) =$

$$\begin{aligned} &= G(x) + \sum_{i=1}^l \frac{g_i(x)}{(v(n))^{i/2}} \int_{1/v(n)}^\infty \frac{1}{y^{i/2}} d\left(H(y - u(n)) + \sum_{j=1}^m \frac{h_j(y - u(n))}{n^{j/2}}\right) = \\ &= G(x) + \sum_{i=1}^l g_i(x) \int_1^\infty z^{-i/2} d\left(H(z/v(n) - u(n)) + \sum_{j=1}^m n^{-j/2} h_j(z/v(n) - u(n))\right). \end{aligned}$$

THE STUDENT ASYMPTOTIC DISTRIBUTION

In [3] it was shown that if the random sample size N_n has the negative binomial distribution with parameters $p = 1/n$ and $r > 0$, that is,

$$\mathbb{P}(N_n = k) = \frac{(k+r-2) \cdot \dots \cdot r}{(k-1)!} \frac{1}{n^r} \left(1 - \frac{1}{n}\right)^{k-1}, \quad k \in \mathbb{N} \quad (3.1)$$

(with $r = 1$ this is the geometric distribution), then for an asymptotically normal statistic T_n we have

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}(\sigma \sqrt{n}(T_{N_n} - \mu) < x) - S_{2r}(x\sqrt{r}) \right| = 0$$

(see [3], Corollary 2.1), where $S_f(x)$ is the d. f. of the Student law with the parameter $f = 2r$ corresponding to the density

$$p_f(x) = \frac{\Gamma(f+1/2)}{\sqrt{\pi f} \Gamma(f/2)} \left(1 + \frac{x^2}{f}\right)^{-(f+1)/2}, \quad x \in \mathbb{R},$$

$\Gamma(\cdot)$ is Euler's gamma-function, and $f > 0$ is the shape parameter (if f is natural-valued, then it is referred as "the number of degrees of freedom"). In general, this parameter can be arbitrarily small corresponding to the case of heavy tails. If $f = 2$, that is $r = 1$, then the d. f. $S_2(x)$ can be expressed explicitly (see the preceding section). With $r = 1/2$ we have the Cauchy distribution.

In the book [4] (see formula (6.112) there on p. 233) the following convergence rate estimate was obtained:

$$\sup_{x \geq 0} \left| \mathbb{P}\left(\frac{N_n}{\mathbb{E}N_n} < x\right) - H_r(x) \right| \leq$$

$$\begin{cases} C_r n^{-1}, & r \geq 1, \\ C_r n^{-r}, & r \in (0, 1), \end{cases} \quad C_r > 0, \quad n \in \mathbb{N}, \quad (3.2)$$

where $H_r(x)$ is the gamma-distribution function with parameter $r > 0$,

$$H_r(x) = \frac{r^r}{\Gamma(r)} \int_0^x e^{-ry} y^{r-1} dy, \quad x \geq 0. \quad (3.3)$$

Furthermore,

$$\mathbb{E}N_n = r(n-1) + 1. \quad (3.4)$$

Thus, from (3.2) – (3.4) it follows that the random sample size N_n satisfies Condition 2.2 with

$$\begin{aligned} g(n) &= r(n-1) + 1, \quad H(x) = H_r(x), \\ m &= 1, \quad h_1(x) \equiv 0, \quad C_2 = C_r > 0, \end{aligned} \quad (3.5)$$

$$\beta = \begin{cases} 1, & r \geq 1, \\ r, & r \in (0, 1). \end{cases} \quad (3.6)$$

Using the equality

$$(1+x)^\gamma = \sum_{k=0}^\infty \frac{\gamma(\gamma-1) \cdot \dots \cdot (\gamma-k+1)}{k!} x^k, \quad |x| < 1, \quad \gamma \in \mathbb{R},$$

it is easy to verify that for $r > 0$, $r \neq 1$, $n \in \mathbb{N}$

$$\mathbb{E}N_n^{-1} = \frac{1 - n^{r-1}}{(n-1)(1-r)n^{r-1}} = O\left(\frac{1}{n^r}\right). \quad (3.7)$$

For $r = 1$ we have

$$\mathbb{E}N_n^{-1} = \frac{1}{n-1} \log n, \quad n > 1. \quad (3.8)$$

So, with the account of Theorem 2.1 we have

$$\begin{aligned} &\int_{1/(r(n-1)+1)}^\infty \Phi(x\sqrt{y}) dH_r(y) = \\ &= \int_0^\infty \Phi(x\sqrt{y}) dH_r(y) + O\left(\frac{1}{n}\right) = S_{2r}(x) + O\left(\frac{1}{n}\right), \quad (3.9) \\ &x|x| \int_{1/(r(n-1)+1)}^\infty \varphi(x\sqrt{y}) \sqrt{y} dH_r(y) = \\ &= x|x| \int_0^\infty \varphi(x\sqrt{y}) \sqrt{y} dH_r(y) + O\left(\frac{1}{n}\right) \equiv \\ &\equiv \frac{x|x| r^r \Gamma(r+1/2)}{\sqrt{2\pi}(r+x^2/2)^{r+1/2} \Gamma(r)} + O\left(\frac{1}{n}\right), \quad (3.10) \end{aligned}$$

Hence, we obtain the following statement.

Theorem 3.1. Let Conditions 1.1 and 1.2 hold and for some $r > \frac{1}{2}$ the random variable N_n has the negative binomial distribution (3.1). Then, as $n \rightarrow \infty$, we have

$$\begin{aligned} &\sup_x \left| \mathbb{P}_\theta(2p_0 \sqrt{2m}(M_{N_n} - \theta) < x) - S_{2r}(x) - \right. \\ &\left. - \frac{p_1 \Gamma(r+1/2) x |x|^r}{2p_0 (r+x^2/2)^{r+1/2} \Gamma(r) \sqrt{2\pi} \sqrt{r(n-1)+1}} \right| = \end{aligned}$$

$$= \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases}$$

where the function $S_{2r}(x)$ was defined in (3.9) and $m = [(r(n-1) + 1)/2]$.

Corollary 3.1. *Let Conditions 1.1 and 1.2 hold and for some $r > \frac{1}{2}$ the random variable N_n has the negative binomial distribution (3.1).*

1. *In the case of Laplace distribution (1.5) for the d. f. of the sample median M_n we have the a. e. of the form*

$$\begin{aligned} & \sup_x \left| \mathbb{P}_\theta(\sqrt{2m}(M_{N_n} - \theta) < x) - S_{2r}(x) - \right. \\ & \left. - \frac{\Gamma(r + 1/2)x|x|^{r-1}}{2\Gamma(r)\sqrt{\pi}(r + x^2/2)^{r+1/2}\sqrt{r(n-1) + 1}} \right| = \\ & = \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases} \end{aligned}$$

where the function $S_{2r}(x)$ was defined in (3.9).

2. *In the case of Cauchy distribution (1.4) we have the a. e.*

$$\begin{aligned} & \sup_x \left| \mathbb{P}_\theta(2\sqrt{2m}(M_{N_n} - \theta) < \pi x) - S_{2r}(x) \right| = \\ & = \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases} \end{aligned}$$

Now define the normalized sample median as

$$\tilde{M}_n = \begin{cases} \sqrt{n-1}X_{(m+1)}, & n = 2m + 1, \\ \frac{1}{2}\sqrt{n}(X_{(m)} + X_{(m+1)}), & n = 2m, m \in \mathbb{N}. \end{cases} \quad (3.11)$$

With the account of the formula

$$\begin{aligned} & \int_{1/(r(n-1)+1)}^\infty \frac{\sqrt{2}p_1x|x|}{4p_0\sqrt{n}} \frac{\varphi(x)}{\sqrt{(r(n-1)+1)}\sqrt{y}} \frac{1}{\sqrt{y}} dH_r(y) \equiv \\ & \equiv \frac{\varphi(x)p_1x|x|\sqrt{2r}}{4p_0\sqrt{rn(n-1)+n}} \frac{\Gamma(r-1/2)}{\Gamma(r)} + O\left(\frac{1}{n}\right), \end{aligned} \quad (3.12)$$

and Theorem 2.2 we obtain the following statement.

Theorem 3.2. *Let Conditions 1.1 and 1.2 hold and for some $r > \frac{1}{2}$ the random variable N_n has the negative binomial distribution (3.1). Then, as $n \rightarrow \infty$, for the d. f. of the sample median we have the a. e.*

$$\begin{aligned} & \sup_x \left| \mathbb{P}_\theta(2p_0(M_{N_n} - \sqrt{2m}\theta) < x) - \Phi(x) - \right. \\ & \left. - \frac{\varphi(x)p_1x|x|\sqrt{2r}}{4p_0\sqrt{rn(n-1)+n}} \frac{\Gamma(r-1/2)}{\Gamma(r)} \right| = \\ & = \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases} \end{aligned}$$

Corollary 3.2. *Let Conditions 1.1 and 1.2 hold and for some $r > \frac{1}{2}$ the random variable N_n has the negative binomial distribution (3.1).*

1. *In the case of Laplace distribution (1.5) we have the following a. e. for the d. f. of the normalized sample median \tilde{M}_n :*

$$\begin{aligned} & \sup_x \left| \mathbb{P}_\theta(\tilde{M}_{N_n} - \sqrt{2m}\theta < x) - \Phi(x) - \right. \\ & \left. - \frac{\varphi(x)x|x|\sqrt{r}}{2\sqrt{rn(n-1)+n}} \frac{\Gamma(r-1/2)}{\Gamma(r)} \right| = \\ & = \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases} \end{aligned}$$

2. *In the case of Cauchy distribution (1.4) we have the following a. e. for the d. f. of the normalized sample median \tilde{M}_n :*

$$\begin{aligned} & \sup_x \left| \mathbb{P}_\theta((\tilde{M}_{N_n} - \sqrt{2m}\theta) < \pi x) - \Phi(x) \right| = \\ & = \begin{cases} O(n^{-1} \log n), & r = 1, \\ O(n^{-1}), & r > \frac{1}{2}, r \neq 1, \end{cases} \end{aligned}$$

THE LAPLACE ASYMPTOTIC DISTRIBUTION

Let $\theta > 0$. Consider the Laplace distribution with the d. f. $\Lambda_\theta(x)$ and density

$$\lambda_\theta(x) = \frac{1}{\theta\sqrt{2}} \exp\left\{-\frac{\sqrt{2}|x|}{\theta}\right\}, \quad x \in \mathbb{R}.$$

In [5] the following example was presented of a sequence of random variables $N_n(s)$ depending on the parameter $s \in \mathbb{N}$. Let Y_1, Y_2, \dots be independent identically distributed random variables with the continuous d. f. For $s \in \mathbb{N}$ define the random variable

$$N(s) = \min\{i \geq 1 : \max_{1 \leq j \leq s} Y_j < \max_{s+1 \leq k \leq s+i} Y_k\}.$$

It is well known that the random variables so defined have the so-called discrete Pareto distribution

$$\mathbb{P}(N(s) \geq k) = \frac{s}{s+k-1}, \quad k \geq 1 \quad (4.1)$$

(see, e. g., [21] or [20]). Now let $N^{(1)}(s), N^{(2)}(s), \dots$ be independent identically distributed random variables with distribution (4.1). Define the random variable

$$N_n(s) = \max_{1 \leq j \leq n} N^{(j)}(s).$$

Then, as it was shown in [5],

$$\lim_{n \rightarrow \infty} \sup_{x > 0} \left| \mathbb{P}(N_n(s) < nx) - e^{-s/x} \right| = 0, \quad (4.2)$$

and for an asymptotically normal statistic T_n we have the relation

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}(\sigma\sqrt{n}(T_{N_n(s)} - \mu) < x) - \Lambda_{1/s}(x) \right| = 0,$$

where $\Lambda_{1/s}(x)$ is the Laplace d. f. with $\theta = 1/s$.

In [19] the following estimate of the rate of convergence in (4.2) was obtained: there exists a constant $C_s \in (0, \infty)$ such that

$$\sup_{x \geq 0} |P(N_n(s) < nx) - e^{-s/x}| \leq C_s n^{-1}, \quad C_s > 0, \quad n \in \mathbb{N}. \quad (4.3)$$

So, from (4.3) it follows that the random variable $N_n(s)$ satisfies Condition 2.2 with

$$g(n) = n, \quad H(x) = e^{-s/x}, \quad m = 1, \quad (4.4)$$

$$h_1(x) \equiv 0, \quad C_2 = C_s > 0, \quad \beta = 1. \quad (4.5)$$

Consider $EN_n^{-1}(s)$ in more detail. From the definition of $N_n(s)$ and (4.1) we obtain

$$\begin{aligned} P(N_n(s) = k) &= \binom{k}{s+k}^n - \binom{k-1}{s+k-1}^n = \\ &= sn \int_{k-1}^k \frac{x^{n-1}}{(s+x)^{n+1}} dx. \end{aligned}$$

Therefore

$$\begin{aligned} EN_n^{-1}(s) &= \sum_{k=1}^{\infty} \frac{P(N_n(s) = k)}{k} = \\ &= sn \sum_{k=1}^{\infty} \frac{1}{k} \int_{k-1}^k \frac{x^{n-1}}{(s+x)^{n+1}} dx \leq \\ &\leq sn \sum_{k=1}^{\infty} \int_{k-1}^k \frac{x^{n-2}}{(s+x)^{n+1}} dx = sn \int_0^{\infty} \frac{x^{n-2}}{(s+x)^{n+1}} dx. \end{aligned}$$

To calculate the last integral use the relation

$$\int_0^{\infty} \frac{x^{s-1}}{(a+bx)^{s+n}} dx = \frac{\Gamma(s)\Gamma(n)}{a^n b^s \Gamma(s+n)}, \quad a, b, s, n > 0,$$

see [11], formula 856.12. We obtain

$$EN_n^{-1}(s) \leq sn \frac{\Gamma(n-1)\Gamma(2)}{s^2 \Gamma(n+1)} = \frac{1}{s(n-1)} = O(n^{-1}).$$

So, with the account of Theorem 1.1 and the formulas

$$\begin{aligned} \int_{n-1}^{\infty} \Phi(x\sqrt{y}) d_y e^{-s/y} &= \int_0^{\infty} \Phi(x\sqrt{y}) d_y e^{-s/y} + O\left(\frac{1}{n}\right) = \\ &= \Lambda_{1/s}(x) + O\left(\frac{1}{n}\right), \end{aligned} \quad (4.5)$$

$$\begin{aligned} x|x| \int_{n-1}^{\infty} \varphi(x\sqrt{y}) \sqrt{y} d_y e^{-s/y} &= \\ = x|x| \int_0^{\infty} \varphi(x\sqrt{y}) \sqrt{y} d_y e^{-s/y} + O\left(\frac{1}{n}\right) &\equiv l_s(x) + O\left(\frac{1}{n}\right), \end{aligned} \quad (4.6)$$

we directly obtain the following theorem.

Theorem 4.1. *Let Conditions 1.1 and 1.2 hold. Assume that for some $s \in \mathbb{N}$ the random variable $N_n(s)$ has the distribution*

$$P(N_n(s) = k) = \binom{k}{s+k}^n - \binom{k-1}{s+k-1}^n, \quad k \in \mathbb{N}. \quad (4.7)$$

Then, as $n \rightarrow \infty$, for the d. f. of the sample median $M_{N_n(s)}$ we have the a. e.

$$P_{\theta}(2p_0\sqrt{2k}(M_{N_n(s)} - \theta) < x) - \Lambda_{1/s}(x) - \frac{p_1 l_s(x)}{2p_0\sqrt{n}} = O\left(\frac{1}{n}\right)$$

uniformly in $x \in \mathbb{R}$, where the functions $\Lambda_{1/s}(x)$ and $l_s(x)$ were defined in (4.5) and (4.6), respectively.

Corollary 4.1. *Let Conditions 1.1 and 1.2 hold. Assume that for some $s \in \mathbb{N}$ the random variable $N_n(s)$ has the distribution (4.7).*

1. As $n \rightarrow \infty$, for Laplace distribution (1.5) we have

$$\sup_x |P_{\theta}(\sqrt{2k}(M_{N_n(s)} - \theta) < x) - \Lambda_{1/s}(x) - \frac{l_s(x)}{\sqrt{2n}}| = O\left(\frac{1}{n}\right),$$

where the functions $\Lambda_{1/s}(x)$ and $l_s(x)$ were defined in (4.5) and (4.6), respectively.

2. As $n \rightarrow \infty$, for Cauchy distribution (1.4) we have

$$\sup_x |P_{\theta}(2\sqrt{2k}(M_{N_n(s)} - \theta) < \pi x) - \Lambda_{1/s}(x)| = O(n^{-1}).$$

For the normalized sample median \tilde{M}_n (see (3.11)), using the formula

$$\int_{n-1}^{\infty} \frac{x|x|}{\sqrt{y}} d_y e^{-s/y} \equiv l_s(x) + O\left(\frac{1}{n}\right),$$

we obtain the following theorem.

Theorem 4.2. *Let Conditions 1.1 and 1.2 hold. Assume that for some $s \in \mathbb{N}$ the random variable $N_n(s)$ has the distribution (4.7). Then, as $n \rightarrow \infty$, for the d. f. of the normalized sample median $\tilde{M}_{N_n(s)}$ we have the a. e.*

$$P_{\theta}(2p_0(\tilde{M}_{N_n(s)} - \sqrt{2m}\theta) < x) - \Phi(x) - \frac{\sqrt{2}p_1 l_s(x)}{4p_0\sqrt{n}} = O\left(\frac{1}{n}\right)$$

uniformly in $x \in \mathbb{R}$, where the function $l_s(x)$ was defined in (4.6).

Corollary 4.2. *Let Conditions 1.1 and 1.2 hold. Assume that for some $s \in \mathbb{N}$ the random variable $N_n(s)$ has the distribution (4.7).*

1. For the Laplace distribution (1.5) we have

$$\sup_x |P_{\theta}(\tilde{M}_{N_n(s)} - \sqrt{2m}\theta) < x) - \Phi(x) - \frac{l_s(x)}{\sqrt{2n}}| = O\left(\frac{1}{n}\right).$$

2. For the Cauchy distribution (1.4) we have

$$\sup_x |P_{\theta}(2(\tilde{M}_{N_n(s)} - \sqrt{2m}\theta) < \pi x) - \Phi(x)| = O(n^{-1}).$$

CONCLUSION

In the paper a general transfer theorem was presented for the asymptotic expansions of the distribution of the sample median constructed from a sample with random size. This theorem gives an algorithm for the construction of these asymptotic expansions from the given asymptotic expansion for the distribution of the sample median in a sample with a non-random size and the given asymptotic expansion for

the distribution of the random sample size. The bounds for the corresponding residuals were also presented in terms of O- and o-symbols. As examples of the application of the general theorem, two special cases were considered where the asymptotic distributions of the sample median in a sample with random size are normal scale mixtures such as the Laplace and Student laws. Moreover, the examples related to samples from the Cauchy, Laplace and Student laws were considered as well. This approach can be successfully used for big data mining and analysis of information flows in high-performance computing.

ACKNOWLEDGEMENTS

Research supported by the Russian Foundation for Basic Research (project 15-07-02652).

REFERENCES

- [1] *V.E. Bening*. Asymptotic Theory of Testing Statistical Hypotheses: Efficient Statistics, Optimality, Power Loss, and Deficiency. – Utrecht: VSP, 2000.
- [2] *V.E. Bening, V.Yu. Korolev*. Generalized Poisson Models and their Applications in Insurance and Finance. – Utrecht: VSP, 2002.
- [3] *V.E. Bening, V.Yu. Korolev*. On an application of the Student distribution in the theory of probability and mathematical statistics // *Theory of Probability and its Applications*, 2005. Vol. 49. No. 3. P. 377–391.
- [4] *V.E. Bening, V.Yu. Korolev, I.A. Sokolov, S.Yu. Shorgin*. Randomized Models and Methods of the Theory of Reliability of Information and Technical Systems. – Moscow: Torus Press, 2007.
- [5] *V.E. Bening, V.Yu. Korolev*. Some statistical problems related to the Laplace distribution // *Informatics and Its Applications*, 2008. Vol. 2. No. 2. P. 19–34.
- [6] *V.E. Bening*. On the deficiency of some estimators based on samples with random sizes // *Bulletin of Tver State University, Series “Applied Mathematics”*, 2015. No. 1. P. 5–14.
- [7] *V.E. Bening, N.K. Galieva, V.Yu. Korolev*. Asymptotic expansions for the distribution functions of statistics constructed from samples with random sizes // *Informatics and Its Applications*, 2013. Vol. 7. No. 2. P. 75–91.
- [8] *V.E. Bening, V.A. Savushkin*. On approximations to the distributions of statistics constructed from samples with random sizes // *Bulletin of Tver State University, Series “Applied Mathematics”*, 2014. No. 1. P. 91–112.
- [9] *M.V. Burnashev*. Asymptotic expansions for the median estimator of the parameter // *Theory Probab. Appl.*, 1996. Vol. 41. No. 4. P. 738–753.
- [10] *H. Cramer*. *Mathematical Methods of Statistics*. – Princeton: Princeton University Press, 1946.
- [11] *H.B. Dwight*. *Tables of Integrals and Other Mathematical Data*, 4th ed. – New York: Macmillan, 1961.
- [12] *B.V. Gnedenko*. On estimation of unknown parameters from a random number of independent observations // *Transactions of Razmadze Tbilisi Mathematical Institute*, 1989. Vol. 92. P. 146–150 (in Russian).
- [13] *B.V. Gnedenko, V.Yu. Korolev*. *Random Summation: Limit Theorems and Applications*. – Boca Raton: CRC Press, 1996.
- [14] *J.L. Hodges, E.L. Lehmann*. Deficiency // *Ann. Math. Statist.*, 1970. Vol. 41. No. 5. P. 783–801.
- [15] *Korolev V. Yu., Chertok A. V., Korchagin A. Yu., Zeifman A. I.* Modeling high-frequency order flow imbalance by functional limit theorems for two-sided risk processes // *Applied Mathematics and Computation*, 2015. Vol. 253. P. 224–241.
- [16] *Korolev V. Yu., Korchagin A. Yu., Zeifman A.I.* Convergence of non-homogeneous random walks generated by compound Cox processes to generalized variance-gamma Levy processes // *Doklady Mathematics*, 2015. Vol. 92. No. 1. P. 408–411.
- [17] *E.L. Lehmann*. *Elements of Large-Sample Theory*. – Berlin–New York: Springer, 1999.
- [18] *E.L. Lehmann, G. Casella*. *Theory of Point Estimation*. 2nd Edition. – Berlin–New York: Springer, 2003.

- [19] *O.O. Lyamin*. On the rate of convergence of the distributions of some statistics to the Laplace and Student distributions // *Bulletin of Moscow State University, Series 15 Computational Mathematics and Cybernetics*, 2011. No. 1. P. 39–47.
- [20] *V.B. Nevzorov*. *Records*. Mathematical Theory. – Moscow: Fakis, 2000.
- [21] *S.S. Wilks*. Recurrence of extreme observations // *Journal of American Mathematical Society*, 1959. Vol. 1, No. 1, P. 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, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. His email is bening@yandex.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, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

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, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences; principal scientist, Institute of Socio-Economic Development of Territories, Russian Academy of Sciences. His email is a_zeifman@mail.ru.

UNIFORM IN TIME BOUNDS FOR “NO-WAIT” PROBABILITY IN QUEUES OF $M_t/M_t/S$ TYPE

Alexander Zeifman
Vologda State University,
Vologda, Russia,
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia
ISEDT RAS

Rostislav Razumchik
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia,
Peoples' Friendship University,
Moscow, Russia

Anna Korotysheva
Yacov Satin
Galina Shilova
Vologda State University,
Vologda, Russia,
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia

Victor Korolev
Moscow State University,
Moscow, Russia
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia

Sergey Shorgin
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia

KEYWORDS

inhomogeneous continuous-time Markov chain, approximation bounds

ABSTRACT

In this paper we present new analytical results concerning long-term staffing problem in high-level telecommunication service systems. We assume that a service system can be modelled either by a classic $M_t/M_t/S$ queue, or $M_t/M_t/S$ queue with batch service or $M_t/M_t/S$ with catastrophes and batch arrivals when empty. The question under consideration is: how many servers guarantee that in the long run the probability of zero delay in a queue is higher than the target probability at all times? Here the methodology is presented, which allows one to construct uniform in time upper bound for the value of S in each of the three cases and does not require the calculation of the limiting distribution. These upper bounds can be easily computed and are accurate enough whenever the arrival intensity is low, but become rougher as the arrival intensity is further increased. In the numerical section one compares the accuracy of the obtained bounds with the exact values of S , obtained by direct numerical computation of the limiting distribution.

1 INTRODUCTION

In this paper consideration is given to three classes of continuous-time Markov chains, which describe the behaviour of multi-server queueing systems of type $M_t/M_t/S$ with a single queue of infinite capacity. Specifically our attention is paid to a classic $M_t/M_t/S$ queue,

$M_t/M_t/S$ queue with batch service, and $M_t/M_t/S$ with catastrophes and batch arrivals when empty. The problem under consideration can be formulated as follows: determine the number of servers S , which guarantees that in the long run the probability of zero waiting time in a queue is higher than the target probability at all times¹.

One can think of two common approaches to the problem. The first one is the numerical. It requires the calculation of the limiting system's state probability distribution (which still depends on time t due to inhomogeneity), by truncating the countable state space of the Markov chain. The truncation must be wise in the sense that the calculation errors must be kept low. Having obtained the limiting probabilities one can find the appropriate value for S by exhaustive search. The second approach is the construction of bounds for the value of S without the calculation of the limiting distribution but by using general inequalities known for some classes of continuous-time Markov chains. The bounds obtained in such a way are not sharp, but do not require the solution of the system of ordinary differential equations. Due to the simplicity of their calculation in some cases (for example, when one needs to know only the order of magnitude of S) they still can be valuable. Moreover even crude bounds facilitate the search of exact values of S , especially when the traffic intensity is high.

In this paper we dwell on the second approach and provide the new methodology which allows one to compute bounds for S in case of three different classes of continuous-time Markov chains introduced above. This methodology heavily relies on the results of Zeifman and Sipin et al. (2015) and thus here one will omit most of the intermediate calculations. It is also worth noticing

¹That is a uniform in time bound.

that the study of qualitative and quantitative properties of inhomogeneous continuous-time Markov chains has received considerable attention since 1980's (one can refer to papers Di Crescenzo et al. (2008)-Zeifman (1995), Zeifman and Korotysheva et al. (2012)-Margoulis (2013) and references therein). Other interesting results related to this paper can be found in Zeifman and Satin et al. (2009) and Zeifman, Satin and Korolev et al. (2014).

The result, which seems to be the most useful for the computation of the bounds for S , can be found in Zeifman and Sipin et al. (2015). According to it, the limiting state probabilities of an inhomogeneous Markov chain $\{X(t), t \geq 0\}$ defined on non-negative integers which described the behaviour of one of the three queueing systems mentioned at the beginning of the Section, satisfy the inequality

$$\limsup_{t \rightarrow \infty} \sum_i d_i p_i(t) \leq K, \quad (1)$$

where K is some constant and $\{d_i\}, i = 0, 1, 2, \dots$ is an increasing sequence of positive numbers, with $d_0 = 1$. As the inequality (1) implies that

$$\limsup_{t \rightarrow \infty} \sum_{i \geq S} d_i p_i(t) \leq \frac{K}{d_S},$$

then the solution of the considered problem can be found from the condition:

$$\Pr(X(t) \leq S) \geq 1 - \frac{K}{d_S}, \quad (2)$$

for sufficiently large t . In the rest of the paper we show how one can obtain in a unified way the values for constants in (2) for each of the chains, mentioned at the beginning of the Section. In order to illustrate the accuracy of the obtained bounds for S we present the comparison of their values with the exact² values of S obtained by applying the first approach (i.e. by truncation of the state space and direct calculation of the limiting probabilities). The details of the direct calculation are omitted and can be found in Zeifman and Korotysheva et al. (2015). From the examples one can see, that the bounds are good as long as the arrival intensity is not too high. As the arrival intensity grows, the bounds become crude.

The problem considered in the paper is closely related to the staffing problem in service systems (such as telephone call centers), in which there is an objective of immediately serving all incoming requests. Specifically, the question is: if one needs to keep, say, more than $X\%$ of incoming requests to be served without waiting, what is the number of service units which allows the system to achieve this goal? There is a big number of research papers devoted to stochastic modelling of service systems in various settings (see, for example, Whitt (2002), Whitt (2002) and references therein) and specifically to the latter question (see, for example, Whitt (2002), Whitt and Song-Hee (2014), Whitt and Liu (2012), Engblom

²In fact they are approximate too, but sharper.

and Pender (2014) and references therein). But in case when the intensities of all the processes (arrivals, services, breakdowns, etc.) vary periodically over time, this question to our knowledge remains open and challenging. The results presented in this paper may give insights into the influence of periodicities on the management of the service system in case when it is possible to model it as $M_t/M_t/S$ queue with possible group services, breakdowns and group arrivals after recovery.

The paper is organized as follows. In the next section description of the general inhomogeneous Markov chain under consideration is given. In Section 3 we show how one can construct the bounds for the value of S for three particular cases of the chain. In the numerical section one demonstrates the accuracy of the obtained results.

2 DESCRIPTION OF THE MODEL

Let $\{X(t), t \geq 0\}$ be an inhomogeneous continuous-time Markov chain describing the evolution of the number of customers in the system which is of $M_t/M_t/S$ type. Denote the state space of $\{X(t), t \geq 0\}$ by $E = \{0, 1, 2, \dots\}$. Throughout the paper we assume that for the transition probabilities it holds that

$$\Pr(X(t+h) = j | X(t) = i) = q_{i,j}(t)h + \alpha_{i,j}(t,h), \quad j \neq i,$$

where all $\alpha_i(t, h) = -\sum_{j \neq i} \alpha_{i,j}(t, h)$ are $o(h)$ uniformly in i , i. e., $\sup_i |\alpha_i(t, h)| = o(h)$. Additionally we assume that all intensity functions are linear combinations of a finite number of locally integrable on $[0, \infty)$ non-negative functions. Let intensity matrix of the chain be $Q(t) = (q_{i,j}(t))$ with $q_{ii}(t) = -\sum_{j \neq i} q_{i,j}(t)$.

Let $a_{ij}(t) = q_{ji}(t)$ for $j \neq i$, then $a_{ii}(t) = -\sum_{j \neq i} a_{ji}(t)$. According to our approach from Zeifman (1995); Zeifman and Leorato et al. (2006) we assume that the intensity matrix is essentially bounded, i. e.

$$|a_{ii}(t)| \leq L < \infty,$$

for almost all $t \geq 0$.

Denote by $p_{ij}(s, t) = \Pr\{X(t) = j | X(s) = i\}$, $i, j \geq 0$, $0 \leq s \leq t$ the transition probability functions of the chain $\{X(t), t \geq 0\}$ and by $p_i(t) = \Pr\{X(t) = i\}$ – the state probabilities. By $\mathbf{p}(t) = (p_0(t), p_1(t), \dots)^T$, $t \geq 0$, denote the column vector of state probabilities.

Probabilistic dynamics of the considered chain $\{X(t), t \geq 0\}$ is given by the forward Kolmogorov system

$$\frac{d\mathbf{p}(t)}{dt} = A(t)\mathbf{p}(t), \quad (3)$$

where $A(t) = Q^T(t)$ is the transposed intensity matrix of the chain. Throughout the paper by $\|\cdot\|$ we denote the l_1 -norm, i. e., $\|\mathbf{x}\| = \sum_i |x_i|$, and $\|B\| = \sup_j \sum_i |b_{ij}|$ for $B = (b_{ij})_{i,j=0}^{\infty}$. Let Ω be the set all stochastic vectors, i. e. l_1 – vectors with nonnegative coordinates and unit norm. Then we have $\|A(t)\| = 2 \sup_k |a_{kk}(t)| \leq 2L$ 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 (3) as a differential equation in the space l_1 with bounded operator.

It is well known (see Dalecki and Krein (1974)) that the Cauchy problem for differential equation (3) has a unique solution for 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 expected value (mean) of the chain $X(t)$ at moment t under initial condition $X(0) = k$.

Recall that 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 (3). chain $X(t)$ has the *limiting mean* $\varphi(t)$, if $\lim_{t \rightarrow \infty} (\varphi(t) - E(t, k)) = 0$ for any k .

3 THEORETICAL BOUNDS

3.1 $M_t/M_t/S$ system

The inhomogeneous continuous-time Markov chain $\{X(t), t \geq 0\}$ describing the behaviour of the ordinary $M_t/M_t/S$ system is of birth-and-death type. Its birth and death intensities are equal to $q_{n,n+1}(t) = \lambda_n(t) = \lambda(t)$ and $q_{n,n-1}(t) = \mu_n(t) = \min(n, S)\mu(t)$, respectively.

Consider an increasing sequence of positive numbers $\{d_i\}$, $i = 1, 2, \dots$, $d_1 = 1$, and the corresponding triangular matrix D :

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

Let l_{1D} be the space of sequences:

$$l_{1D} = \{\mathbf{z} = (p_1, p_2, \dots)^T : \|\mathbf{z}\|_{1D} \equiv \|D\mathbf{z}\| < \infty\}.$$

Put

$$d = \inf_{i \geq 1} d_i = 1, \quad W = \inf_{i \geq 1} \frac{d_i}{i}, \quad g_i = \sum_{n=1}^i d_n.$$

Consider the following expressions:

$$\alpha_k(t) = \lambda_k(t) + \mu_{k+1}(t) - \frac{d_{k+1}}{d_k} \lambda_{k+1}(t) - \frac{d_{k-1}}{d_k} \mu_k(t), \quad k \geq 0, \quad (4)$$

and

$$\alpha(t) = \inf_{k \geq 0} \alpha_k(t). \quad (5)$$

The property $\mathbf{p}(t) \in \Omega$ for any $t \geq 0$ allows one to use the normalization condition and write $p_0(t) = 1 - \sum_{i \geq 1} p_i(t)$. Then from (3) we obtain the following system of differential equations for the considered chain $\{X(t), t \geq 0\}$:

$$\frac{d\mathbf{z}(t)}{dt} = B(t)\mathbf{z}(t) + \mathbf{f}(t), \quad (6)$$

where $\mathbf{z}(t) = (p_1(t), p_2(t), \dots)^T$, $\mathbf{f}(t) = (\lambda_0(t), 0, 0, \dots)^T$, $B(t) = (b_{ij}(t))_{i,j=1}^{\infty}$ and

$$b_{ij} = \begin{cases} -(\lambda_0 + \lambda_1 + \mu_1), & \text{if } i = j = 1, \\ \mu_2 - \lambda_0, & \text{if } i = 1, j = 2, \\ -\lambda_0, & \text{if } i = 1, j > 2, \\ -(\lambda_j + \mu_j), & \text{if } i = j > 1, \\ \mu_j, & \text{if } i = j - 1 > 1, \\ \lambda_j, & \text{if } i = j + 1 > 1, \\ 0, & \text{otherwise.} \end{cases}$$

This is a system of linear non-homogeneous differential equations which solution can be written in the form

$$\mathbf{z}(t) = V(t, 0)\mathbf{z}(0) + \int_0^t V(t, \tau)\mathbf{f}(\tau) d\tau,$$

where $V(t, \tau) = V(t)V^{-1}(\tau)$ is the Cauchy operator of (6).

Consider equation (6) in the space l_{1D} . We have

$$DBD^{-1} =$$

$$\begin{pmatrix} -(\lambda_0 + \mu_1) & \frac{d_1}{d_2}\mu_2 & 0 & \cdots \\ \frac{d_2}{d_1}\lambda_1 & -(\lambda_1 + \mu_2) & \frac{d_2}{d_3}\mu_3 & 0 & \cdots \\ 0 & \frac{d_3}{d_2}\lambda_2 & -(\lambda_2 + \mu_3) & \frac{d_3}{d_4}\mu_4 & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \quad (7)$$

Note that $\mathbf{f}(t)$ and $B(t)$ are bounded and locally integrable on $[0, \infty)$ as being a vector function and an operator function in l_{1D} respectively. Now, taking into consideration (5), we have the following bound for the logarithmic norm $\gamma(B(t))$ in l_{1D} :

$$\begin{aligned} \gamma(B)_{1D} &= \gamma(DB(t)D^{-1})_1 = \\ \sup_{i \geq 0} \left(\frac{d_{i+1}}{d_i} \lambda_{i+1}(t) + \frac{d_{i-1}}{d_i} \mu_i(t) - (\lambda_i(t) + \mu_{i+1}(t)) \right) &= \\ - \inf_{k \geq 0} (\alpha_k(t)) &= -\alpha(t), \end{aligned}$$

Hence

$$\|V(t, s)\|_{1D} \leq e^{-\int_s^t \alpha(\tau) d\tau},$$

and therefore

$$\begin{aligned} \|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} &= \|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\|_{1D} \leq \\ &\leq e^{-\int_s^t \alpha(\tau) d\tau} \|\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\|_{1D}, \end{aligned}$$

for any $t \geq s \geq 0$ and any initial conditions $\mathbf{p}^*(s), \mathbf{p}^{**}(s)$.

Moreover, inequality $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \leq 2\|\mathbf{z}^*(t) - \mathbf{z}^{**}(t)\| \leq 4\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D}$ implies the following bound:

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \leq 4e^{-\int_s^t \alpha(\tau) d\tau} \sum_{i \geq 1} g_i |p_i^*(s) - p_i^{**}(s)|,$$

Assume that the following bounds hold

$$e^{-\int_t^\tau \alpha(u)du} \leq M^* e^{-\alpha^*(t-\tau)}, \quad \lambda(t) \leq \Lambda, \quad (8)$$

for almost all $t \geq 0$. Then we obtain the following inequality

$$\begin{aligned} \|\mathbf{z}(t)\|_{1D} &\leq \\ \|V(t)\|_{1D}\|\mathbf{z}(0)\|_{1D} + \int_0^t \|V(t,\tau)\|_{1D}\|\mathbf{f}(\tau)\|_{1D} d\tau &\leq \\ M^* e^{-\alpha^* t}\|\mathbf{z}(0)\|_{1D} + \frac{M^* \Lambda}{\alpha^*}. \end{aligned}$$

On the other hand, because all $p_i(t)$ are non-negative, we have

$$\|\mathbf{z}(t)\|_{1D} = \sum_{i \geq 1} p_i(t) \sum_{k=1}^i d_k \geq \sum_{i \geq N} g_i p_i(t).$$

Hence

$$\sum_{i=N}^{\infty} g_i p_i(t) \leq M^* e^{-\alpha^* t} \|\mathbf{z}(0)\|_{1D} + \frac{M^* \Lambda}{\alpha^*},$$

and

$$\sum_{i=N}^{\infty} p_i(t) \leq g_N^{-1} M^* e^{-\alpha^* t} \|\mathbf{z}(0)\|_{1D} + \frac{M^* \Lambda}{\alpha^* g_N},$$

for any N and any $t \geq 0$, and we obtain the following theorem.

Theorem 1 Assume that the arrival and service intensities $\lambda(t)$ and $\mu(t)$ for $M_t/M_t/S$ system are known and satisfy (8). Then the chain $\{X(t), t \geq 0\}$, describing the number of customers in the system at time t , is exponentially weakly ergodic, and the following bound holds for any N :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < N) \geq 1 - \frac{M^* \Lambda}{\alpha^* g_N}.$$

From the previous theorem it follows that the inequality (3) for the $M_t/M_t/S$ system has the form

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{M^* \Lambda}{\alpha^* g_S}.$$

3.2 $M_t/M_t/S$ system with group services

The inhomogeneous continuous-time Markov chain $\{X(t), t \geq 0\}$ describing the behaviour of the ordinary $M_t/M_t/S$ system has non-zero arrival intensities $q_{n,n+1}(t) = \lambda(t)$, and non-zero service intensities $q_{n,n-i}(t) = \mu_i(t) = \frac{\mu(t)}{i}$ for group of i customers, $1 \leq i \leq S$. The chain $\{X(t), t \geq 0\}$ is a SZK chain described in detail in Satin and Zeifman et al. (2013); Zeifman and Satin et al. (2013); Zeifman and Korotysheva et al. (2015).

The analysis of the limiting behaviour for this chain can be carried out in the same way as it was done in the previous subsection. We have

$$\alpha_i(t) = -a_{ii}(t) - \sum_{k \geq 1} \lambda_k(t) \frac{d_{k+i}}{d_i} - \sum_{k=1}^{i-1} (\mu_{i-k}(t) - \mu_i(t)) \frac{d_k}{d_i}, \quad (9)$$

and

$$\alpha(t) = \inf_{i \geq 1} \alpha_i(t), \quad (10)$$

instead of (4) and (5), where $\lambda_1(t) = \lambda(t)$, $\lambda_k(t) \equiv 0$ for $k \geq 1$, and $\mu_k(t) = \frac{\mu(t)}{k}$, if $k \leq S$, $\mu_k(t) \equiv 0$ for $k > S$. Instead of (7) we obtain

$$DBD^{-1} = \begin{pmatrix} a_{11} & (\mu_1 - \mu_2) \frac{d_1}{d_2} & (\mu_2 - \mu_3) \frac{d_1}{d_3} & \cdots & (\mu_{r-1} - \mu_r) \frac{d_1}{d_r} & \cdots \\ \lambda_1 \frac{d_2}{d_1} & a_{22} & (\mu_1 - \mu_3) \frac{d_2}{d_3} & \cdots & (\mu_{r-2} - \mu_r) \frac{d_2}{d_r} & \cdots \\ \lambda_2 \frac{d_3}{d_1} & \lambda_1 \frac{d_3}{d_2} & a_{33} & \cdots & (\mu_{r-3} - \mu_r) \frac{d_3}{d_r} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_{r-1} \frac{d_r}{d_1} & \lambda_{r-2} \frac{d_r}{d_2} & \lambda_{r-3} \frac{d_r}{d_3} & \cdots & a_{rr} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Hence the following theorem holds.

Theorem 2 Assume that the arrival and service intensities $\lambda(t)$ and $\mu(t)$ for $M_t/M_t/S$ system with group services are known and assume that (8), (9), (10) hold. Then the chain $\{X(t), t \geq 0\}$, describing the number of customers in the system at time t , is exponentially weakly ergodic, and the following bound holds for any N :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < N) \geq 1 - \frac{M^* \Lambda}{\alpha^* g_N}.$$

From the previous theorem it follows that the inequality (3) for the $M_t/M_t/S$ system with group services has the form

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{M^* \Lambda}{\alpha^* g_S}.$$

3.3 $M_t/M_t/S$ system with catastrophes and batch arrivals when empty

Assume that the inhomogeneous continuous-time Markov chain $\{X(t), t \geq 0\}$ describing the behaviour of the ordinary $M_t/M_t/S$ system with catastrophes and batch arrivals when empty has the following intensity matrix:

$$Q(t) = \begin{pmatrix} a_{00}(t) & r_1(t) & r_2(t) & r_3(t) & r_4(t) & \cdots & \cdots \\ \beta_1(t) + \mu_1(t) & a_{11}(t) & \lambda_1(t) & 0 & 0 & \cdots & \cdots \\ \beta_2(t) & \mu_2(t) & a_{22}(t) & \lambda_2(t) & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \beta_r(t) & 0 & \cdots & \mu_r(t) & a_{rr}(t) & \lambda_r(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

For the detailed description of the system one can refer to Zeifman, Korotysheva and Satin et al. (2015); Zeifman and Satin et al. (2016) and references therein. We will apply the approach from Zeifman, Korotysheva and Satin et al. (2015) or study of ergodic properties and for the construction of bounds for $\{X(t), t \geq 0\}$.

Put

$$\beta_*(t) = \inf_i \beta_i(t),$$

and rewrite the forward Kolmogorov equation (3) as

$$\frac{d\mathbf{p}(t)}{dt} = A^*(t) \mathbf{p}(t) + \mathbf{g}(t), \quad t \geq 0, \quad (11)$$

where $\mathbf{g}(t) = (\beta_*(t), 0, 0, \dots)^T$, $A^*(t) = \{a_{ij}^*(t)\}$, and

$$a_{ij}^*(t) = \begin{cases} a_{0j}(t) - \beta_*(t) & \text{if } i = 0 \\ a_{ij}(t) & \text{otherwise.} \end{cases}$$

Let now D be a diagonal matrix

$$D = \text{diag}(d_0, d_1, d_2, \dots)$$

with elements satisfying the inequalities $1 = d_0 \leq d_1 \leq d_2 \leq \dots$. Consider the corresponding space of sequences $l_{1D} = \{\mathbf{z} = (p_0, p_1, p_2, \dots)\}$ such that $\|\mathbf{z}\|_{1D} = \|D\mathbf{z}\|_1 < \infty$.

Put $\beta_{**}(t) = \inf_i \left(|a_{ii}^*(t)| - \sum_{j \neq i} \left| \frac{d_j}{d_i} a_{ji}^*(t) \right| \right)$.

Then one can obtain the following estimate for the logarithmic norm of the operator function $A^*(t)$ in the l_{1D} -norm:

$$\begin{aligned} \gamma(A^*(t))_{1D} &= \gamma(DA^*(t)D^{-1}) = \\ &= \sup_i \left(a_{ii}^*(t) + \sum_{j \neq i} \left| \frac{d_j}{d_i} a_{ji}^*(t) \right| \right) = -\beta_{**}(t). \end{aligned}$$

Assume now that the following bounds hold:

$$e^{-\int_\tau^t \beta_{**}(u) du} \leq M e^{-a^*(t-\tau)}, \quad (12)$$

$$\beta^*(t) \leq \Theta, \quad (13)$$

for almost all $0 \leq \tau \leq t$. Then one can bound the solution of the system (11) in the following way:

$$\begin{aligned} \|\mathbf{p}(t)\|_{1D} &= \|U^*(t, 0) \mathbf{p}(0) + \int_0^t U^*(t, \tau) \mathbf{g}(\tau) d\tau\| \leq \\ &\leq \|U^*(t, 0)\|_{1D} \|\mathbf{p}(0)\|_{1D} + \int_0^t \|U^*(t, \tau)\|_{1D} \|\mathbf{g}(\tau)\|_{1D} d\tau \leq \\ &\leq M e^{-a^* t} \|\mathbf{p}(0)\|_{1D} + \frac{M\Theta}{a^*}. \end{aligned}$$

On the other hand,

$$\sum_{i \geq N} d_i p_i(t) \leq \|\mathbf{p}(t)\|_{1D},$$

and hence

$$\sum_{i=N}^{\infty} d_i p_i(t) \leq M e^{-a^* t} \|\mathbf{p}(0)\|_{1D} + \frac{M\Theta}{a^*},$$

and

$$\sum_{i=N}^{\infty} p_i(t) \leq d_N^{-1} M e^{-a^* t} \|\mathbf{p}(0)\|_{1D} + \frac{M\Theta}{a^* d_N},$$

for any N and any $t \geq 0$. Thus the following theorem holds.

Theorem 3 Assume that the intensities $\lambda_i(t)$, $\mu_i(t)$, $\beta_i(t)$ and $r_i(t)$ for $M_i/M_t/S$ with catastrophes and batch arrivals when empty are known. Assume that (12) and (13) hold. Then the chain $\{X(t), t \geq 0\}$, describing the number of customers in the system at time t , is exponentially weakly ergodic, and the following bound holds for any N :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < N) \geq 1 - \frac{M\Theta}{a^* d_N}.$$

From the previous theorem it follows that the inequality (3) for the $M_t/M_t/S$ system with catastrophes and batch arrivals when empty has the form

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{M\Theta}{a^* d_S}.$$

4 NUMERICAL EXAMPLES

In this section we will consider six different examples of inhomogeneous Markov chains describing the behaviour of $M_t/M_t/S$ queues. For each example we give two values of the required number of servers S . One of the values is obtained using theorems from the previous section, and the other value is obtained by direct numerical computation of the limiting probabilities (for more details how the limiting probabilities are found one can refer to Zeifman, Korotysheva and Satin et al. (2015); Zeifman and Korotysheva et al. (2015)). As one will see the accuracy of bounds greatly depends on the arrival intensity and drops significantly as it grows.

Example 1

Consider the classic $M_t/M_t/S$ queue with arrival intensity $\lambda(t) = 1 + \sin 2\pi t$ and service intensity $\mu(t) = 3 + \cos 2\pi t$.

Put $d_{n+1} = 2^n$, $n \geq 1$. Then using results from the subsection 3.1 we have that $g_k = 2^k - 1$ and $\alpha_k(t) \geq \mu(t) - \lambda(t)$ for any $k \geq 1$, S and $t \geq 0$. From (5) it follows that $\alpha(t) \geq 2 + \cos 2\pi t - \sin 2\pi t$. Notice that the inequality (8) holds for $\alpha^* = 2$, $M^* = 2$ and $\Lambda = 2$. Hence the Theorem 1 gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{2}{2^S - 1}. \quad (14)$$

One can see that $S = 4$ guarantees that more than 90% of incoming requests will be served without waiting. On the other hand, if one directly computes the limiting probabilities (by truncation of the state space of the chain),

then one can find that in fact $S = 2$ guarantees that 90% of requests have zero waiting time. The behaviour of the probabilities of zero waiting time for $S = 1$ and $S = 2$ is depicted in Fig. 1 and 2.

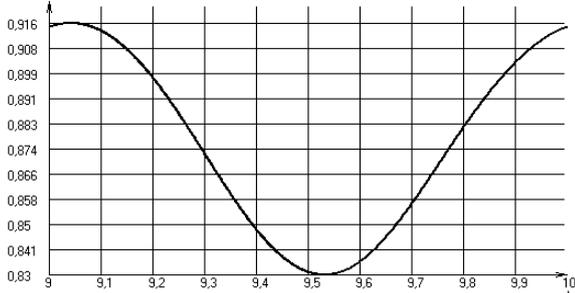


Figure 1: Probability of immediately serving of 80% incoming requests $\Pr(X(t) < 1)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 100$.

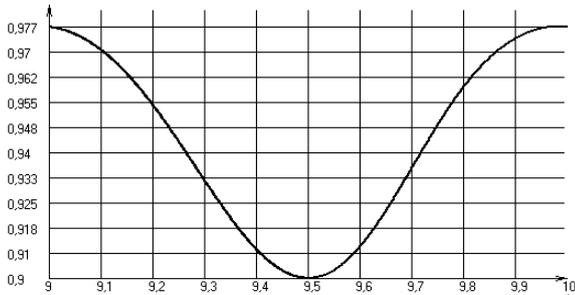


Figure 2: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 2)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 100$.

Example 2

Consider the $M_t/M_t/S$ queue with group services with the arrival intensity $\lambda(t) = 1 + \sin 2\pi t$ and service intensity $\mu(t) = 3 + \cos 2\pi t$. This is the same example as in Zeifman and Satin et al. (2013).

Put $d_{n+1} = 2^n$, $n \geq 1$. Then using results from the subsection 3.2 we have that $g_k = 2^k - 1$ and $\alpha_k(t) \geq \mu(t) - \lambda(t)$, for any $k \geq 1$, S and $t \geq 0$. Therefore from (10) it follows that $\alpha(t) \geq 2 + \cos 2\pi t - \sin 2\pi t$, and the inequality (8) holds for $\alpha^* = 2$, $M^* = 2$ and $\Lambda = 2$. Hence the *Theorem 2* gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{2}{2^S - 1}. \quad (15)$$

It can be seen that even with group services, four servers ($S = 4$) guarantee that more than 90% of incoming requests will be served without waiting. But by numerical computation of the limiting probabilities, one can find that in fact $S = 3$ guarantees that 90% of requests have zero waiting time. The behaviour of the probabilities of

zero waiting time for $S = 2$ and $S = 3$ is depicted in Fig. 3 and 4.

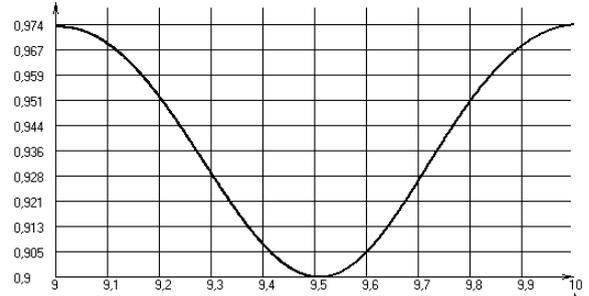


Figure 3: Probability of immediately serving of 87% incoming requests $\Pr(X(t) < 2)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 100$.

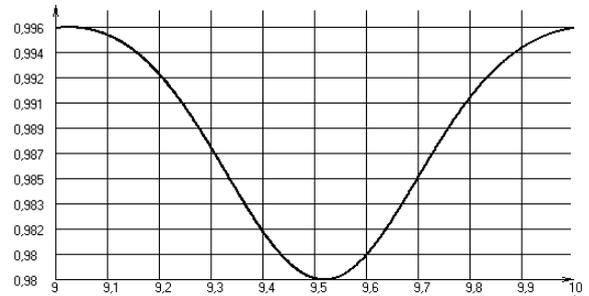


Figure 4: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 3)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 100$.

Example 3

Consider the classic $M_t/M_t/S$ queue service intensity $\mu(t) = 4 + \cos 2\pi t$ but with very high arrival intensity $\lambda(t) = 100 + 5 \sin 2\pi t$. In order to choose the sequence of positive numbers $\{d_i\}$ (see subsection 3.1) we initially suppose that $S \geq 60$. The if one puts $d_{k+1} = 1.02^k$ for $k \geq 1$ then using results from the subsection 3.1 we have that $g_k > 1.02^k$, $k \geq 2$. After little algebra one can find that the inequality (8) is satisfied for $\alpha^* = 2$, $M^* \leq e^2$ and $\Lambda \approx 100$. Hence the *Theorem 1* gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{10^3}{2 \cdot 1.02^S}, \quad (16)$$

Here $S \geq 450$ guarantees that more than 90% of incoming requests will be served without waiting. On the other hand, if one directly computes the limiting probabilities, then one can find that the true value is $S = 37$. The behaviour of the probabilities of zero waiting time for $S = 36$ and $S = 37$ is depicted in Fig. 5 and 6.

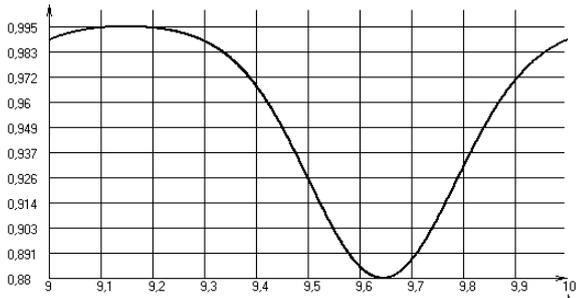


Figure 5: Probability of immediately serving of 85% incoming requests $\Pr(X(t) < 36)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

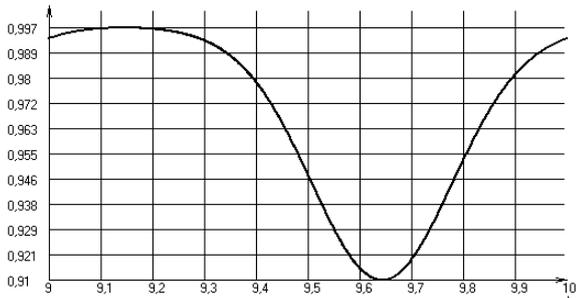


Figure 6: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 37)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

Example 4

Consider the $M_t/M_t/S$ queue with group services with the service intensity $\mu(t) = 3 + \cos 2\pi t$ and high arrival intensity $\lambda(t) = 1 + \sin 2\pi t$. Again in order to choose the sequence of positive numbers $\{d_i\}$ (see subsection 3.2) we initially suppose that $S \geq 60$. The if one puts $d_{k+1} = 1.02^k$ for $k \geq 1$ then using results from subsection 3.2 we have that $g_k > 1.02^k$, $k \geq 2$ and the inequality (8) is satisfied for $a^* = 2$, $M^* \leq e^2$ and $\Lambda \approx 100$. Hence the *Theorem 2* gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{10^3}{2 \cdot 1.02^S}. \quad (17)$$

Surprisingly, as in Example 3, here $S \geq 450$ guarantees that more than 90% of incoming requests will be served without waiting. But the true value is $S = 53$. The behaviour of the probabilities of zero waiting time for $S = 52$ and $S = 53$ is depicted in Fig. 7 and 8.

Example 5

Consider the $M_t/M_t/S$ queue with breakdowns and batch arrivals when empty. Let the arrival and service rates be as in Examples 1,2. Assume that the breakdown and batch arrival intensities are state-dependent and equal to $\beta_n(t) = 2 + \cos 2\pi t + \frac{1}{n}$ and $r_n(t) = \frac{1 - \sin 2\pi t}{4^n}$, $n \geq 1$,

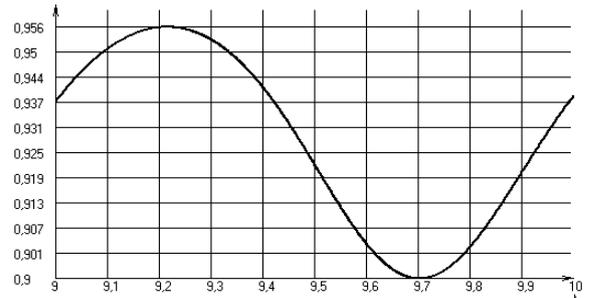


Figure 7: Probability of immediately serving of 88% incoming requests $\Pr(X(t) < 52)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

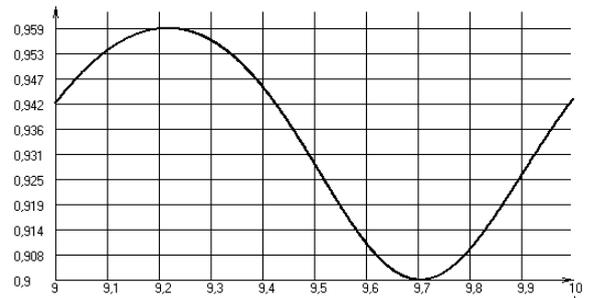


Figure 8: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 53)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

respectively. Notice that breakdowns may happen only when system is not empty.

Put $\beta^*(t) = 1$ and $d_k = \left(\frac{4}{3}\right)^k$ for $k \geq 1$. Then using results from the subsection 3.3 we have that $\beta^{**}(t) \geq \frac{1}{3}$ and (12), (13) hold for $M = 1$, $a^* = \frac{1}{3}$ and $\Theta = 1$. Hence the *Theorem 3* gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{3^{S+1}}{4^S}. \quad (18)$$

One can see that $S \geq 12$ guarantees that more than 90% of incoming requests will be served without waiting. It turns out that this bound is very crude. Indeed from the direct computation of the limiting distribution it follows that in fact already one server ($S = 1$) is enough to guarantee 90% of no-wait. (see Fig. 9).

Example 6

Consider again the $M_t/M_t/S$ queue with breakdowns and batch arrivals when empty but with arrival and service intensities as in Example 3. Assume that the intensities of breakdowns and batch arrivals are $\beta_n(t) = 2 + \cos 2\pi t + \frac{1}{n}$ and $r_n(t) = \frac{1 - \sin 2\pi t}{4^n}$, $n \geq 1$, respectively. Again breakdowns may happen only when system is not empty.

Put $\beta^*(t) = 1$, and $d_k = \left(\frac{201}{200}\right)^k$ for $k \geq 1$. Then using results from the subsection 3.3 we have that $\beta^{**}(t) \geq \frac{2}{5}$ and (12), (13) hold for $M = 1$, $a^* = \frac{2}{5}$ and $\Theta = 1$. Hence

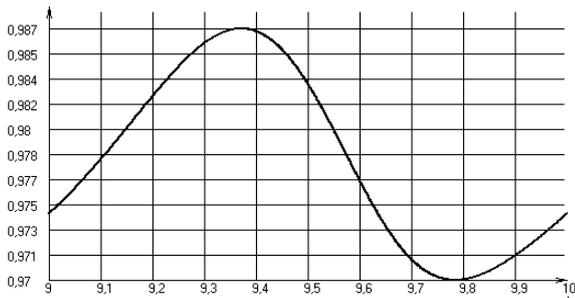


Figure 9: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 1)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 100$.

the *Theorem 3* gives us the following bound for the value of S :

$$\limsup_{t \rightarrow \infty} \Pr(X(t) < S) \geq 1 - \frac{5 \cdot 200^S}{2 \cdot 201^S}. \quad (19)$$

One can see that $S \geq 650$ guarantees more than 90% of incoming requests will be served without waiting. But from numerical computation of limiting probabilities it follows that in fact $S = 13$ is enough (see Fig. 10 and 11).

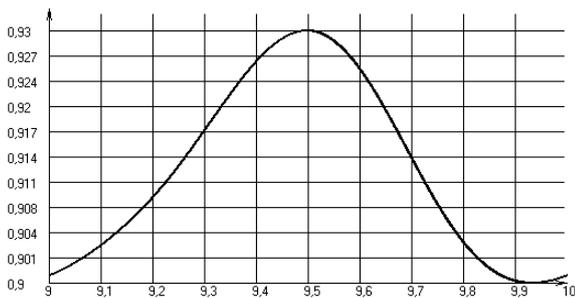


Figure 10: Probability of immediately serving of 88% incoming requests $\Pr(X(t) < 12)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

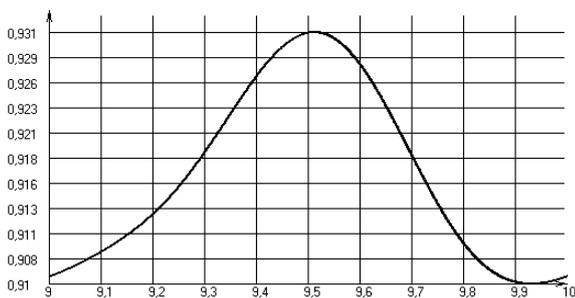


Figure 11: Probability of immediately serving of 90% incoming requests $\Pr(X(t) < 13)$ on $[9, 10]$, with error less than 10^{-6} , one can choose $n = 600$.

5 CONCLUSION

As one can see from the examples the (upper) bounds given by the theorems remain good as long as the ar-

rival intensity is low. As the latter grows the bounds become very inaccurate and additional analysis is needed. Clearly two directions of further research are visible: elaboration of similar bounds for more complex systems and the development of methodology for the estimation of bounds for probabilities given by (3) in cases when the arrival rates are high.

Acknowledgement

This work was supported by Russian Scientific Foundation (Grant No. 14-11-00397).

REFERENCES

- Whitt, W., Jennings, B., Mandelbaum, A., Massey, W. 1996. Server Staffing to Meet Time-Varying Demand. *Management Science*. Vol. 42. No. 10. Pp. 1383-1394.
- Whitt, W. 2006. Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. *Production and Operations Management*. Vol. 15. No. 1. Pp. 88–102.
- Whitt, W. 2002. Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions. Working paper.
- Whitt, W., Song-Hee, K. 2014. Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? *Manufacturing and Service Operations Management*. Vol. 16. No. 3. Pp. 464–480.
- Whitt, W., Liu, Y. 2012. ‘Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research*. Vol. 60. No. 6. Pp. 1551–1564.
- Dalecki, Ju., Krein, M. Stability of solutions of differential equations in Banach space. *American Mathematical Society Translations*. Vol. 43. 386 p.
- Di Crescenzo, A., Giorno, V., Nobile, A., Ricciardi, L. 2008. A note on birth-death processes with catastrophes. *Statistical Probability Letters*. Vol. 78. Pp. 2248–2257.
- Dudin, A., Karolik, A. 2001. BMAP/SM/1 queue with Markovian input of disasters and non-instantaneous recovery. *Performance Evaluation*. Vol. 45. Pp. 19–32.
- Dudin, A., Nishimura, S. 1999. A BMAP/SM/1 queueing system with Markovian arrival input of disasters. *Journal of Applied Probability*. Vol. 36. Pp. 868–881.
- Dudin, A., Semenova, O. 2004. Stable algorithm for stationary distribution calculation for a BMAP/SM/1 queueing system with markovian input of disasters. *Journal of Applied Probability*. Vol. 42. Pp. 547–556.
- Satin, Ya., Zeifman, A., Korotysheva, A. 2013. On the rate of convergence and truncations for a class of Markovian queueing systems. *Theory. Prob. Appl.* Vo. 57. Pp. 529C-539.
- Zeifman A. 1995. Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. *Stochastic Processes and Applications*. Vol. 59. Pp. 157–173.

- Zeifman, A., Leorato, S., Orsingher, E., Satin, Ya., Shilova, G. 2006. Some universal limits for nonhomogeneous birth and death processes. *Queueing Systems*. Vol. 52. Pp. 139–151.
- Zeifman, A., Satin, Ya., Korolev, V., Shorgin, S. 2014. On truncations for weakly ergodic inhomogeneous birth and death processes. *International Journal of Applied Mathematics and Computer Science*. Vol. 24. No. 3. Pp. 503–518.
- Zeifman, A., Korotysheva, A. 2012. Perturbation bounds for $M_t|M_t|N$ queue with catastrophes. *Stochastic models*. Vol. 28. Pp. 49–62.
- Zeifman, A., Satin, Ya., Korotysheva, A., Tereshina N. 2009. On the limiting characteristics for $M(t)|M(t)|S$ queue with catastrophes. *Informatics and its Applications*. Vol. 3. No. 3. Pp. 16–22. (in Russian)
- Zeifman, A., Satin, Ya., Shilova, G., Korolev, V., Bening, V., Shorgin, S. 2013. On $M_t|M_t|S$ type queue with group services. *Proceedings of the 27th European Conference on Modelling and Simulation*. Pp. 604–609.
- Zeifman, A., Korotysheva, A., Satin, Ya., Korolev, V., Bening, V. 2014. Perturbation bounds and truncations for a class of Markovian queues. *Queueing Systems*. Vol. 76. Pp. 205–221.
- Zeifman, A., Korotysheva, A., Satin, Ya., Korolev, V., Shorgin, S., Razumchik, R. 2015. Ergodicity and perturbation bounds for inhomogeneous birth and death processes with additional transitions from and to origin. *Int. J. Appl. Math. Comput. Sci.* Vol. 25. No. 4. Pp. 787–802.
- Zeifman, A., Sipin, A., Korolev, V., Bening, V. 2015. Estimates of some characteristics of multidimensional birth-and-death processes. *Doklady Mathematics*. Vol. 92. Pp. 695–697.
- Zeifman, A., Satin, Ya., Korotysheva, A., Korolev, V., Bening, V. 2016. On a class of Markovian queues with particularities in zero. *Doklady Mathematics*. Submitted.
- Zeifman, A., Korotysheva, A., Razumchik, R., Korolev, V., Shorgin, S. 2015. Some Results For Inhomogeneous Birth-And-Death Process With Application To Staffing Problem In Telecommunication Service Systems. *Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*. Pp. 243–246.
- Engblom, S., Pender, J. 2014. Approximations for the Moments of Nonstationary and State Dependent Birth-Death Queues. eprint arXiv:1406.6164.
- Margoulis, B. 2007. Transient and periodic solution to the time-inhomogeneous quasi-birth death process. *Queueing Systems*. Vol. 56. No. 3. Pp. 183–194.

AUTHOR BIOGRAPHIES

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, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; principal

scientist, Institute of Socio-Economic Development of Territories, Russian Academy of Sciences. His email is a_zeifman@mail.ru and his personal webpage at <http://uni-vologda.ac.ru/zai/eng.html>.

YACOV SATIN is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. His email is yacovi@mail.ru.

ANNA KOROTYSHEVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Vologda State University. Her email is a_korotysheva@mail.ru.

GALINA SHILOVA is Candidate of Science (PhD) in physics and mathematics, associate professor, Head of Department of Mathematics, 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, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. His email is victoryukorolev@yandex.ru.

SERGEY SHORGIN is Doctor of Science in physics and mathematics, professor, Deputy Director of the Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. His email is sshorgin@ipiran.ru.

ROSTISLAV V. RAZUMCHIK received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as a senior research fellow at Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS). Currently he holds the position of Head of the Information and Telecommunication System Modelling section at the FRC CSC RAS and associate professor position at Peoples' Friendship University of Russia. His current research activities focus on queueing theory and its applications in performance evaluation of stochastic systems. His email address is rrazumchik@ipiran.ru.

HYBRID SIMULATION OF ACTIVE TRAFFIC MANAGEMENT

Anna V. Korolkova, Tatyana R. Velieva, Pavel O. Abaev

Department of Applied Probability and Informatics

Peoples' Friendship University of Russia

Miklukho-Maklaya str. 6, Moscow, 117198, Russia

Email: akorolkova@sci.pfu.edu.ru, trvelieva@gmail.com, pabaev@sci.pfu.edu.ru

Leonid A. Sevastianov

Department of Applied Probability and Informatics

Peoples' Friendship University of Russia

Miklukho-Maklaya str. 6, Moscow, 117198, Russia

and Bogoliubov Laboratory of Theoretical Physics

Joint Institute for Nuclear Research

Joliot-Curie 6, Dubna, Moscow region, 141980, Russia

Email: leonid.sevast@gmail.com

Dmitry S. Kulyabov

Department of Applied Probability and Informatics

Peoples' Friendship University of Russia

Miklukho-Maklaya str. 6, Moscow, 117198, Russia

and Laboratory of Information Technologies

Joint Institute for Nuclear Research

Joliot-Curie 6, Dubna, Moscow region, 141980, Russia

Email: yamadharm@gmail.com

KEYWORDS

Hybrid modeling, fluid model, active queue management, random early detection, Modelica

ABSTRACT

For the study and verification of our mathematical model of RED-like active traffic management module a discrete simulation model and a continuous analytical model were developed. However, for various reasons, these implementations are not entirely satisfactory. It is necessary to develop a more adequate simulation model, possibly using a different modeling paradigm. In order to modeling of the TCP source, the RED control module, and the process of their interaction it is proposed to use a hybrid (continuous-discrete) approach. For computer implementation of the model the physical modeling language Modelica is used. Because the language Modelica has multiple implementations we have selected the OpenModelica compiler. The hybrid approach allows us to take into account the transitions between different states in the continuous model of the TCP protocol. The hybrid approach simplified the consideration of the model due to the conversion of a differential inclusions into a set of differential equations with discrete transitions. The considered approach allowed to obtain a simple simulation model of interaction between RED module and TCP source. This model has great potential for expansion. It is possible to implement different types of TCP and RED.

Furthermore, it is possible to use a hybrid approach not only for the simulation but also for analytical modeling.

INTRODUCTION

While complex systems modeling there is the problem of choosing a model approach. When using a discrete-continuous dichotomy, we always have some inappropriate elements. These elements of the model do not well correspond to the selected approach. In particular, the existing models of control systems can not fully meet our needs.

As the implementation of the system with a threshold control we investigated the RED (see Floyd and Jacobson (1993); Feng et al. (2015); Lautenschlaeger and Francini (2015); Karmeshu et al. (2016)) active traffic management unit for the TCP protocol. During simulation of TCP protocol and RED mechanism the serious problems arised. As it turned out adequate TCP models are simply missing. Not even a common method for its modeling (see Paxson and Floyd (1997, 1995); Leland et al. (1994)) exists.

For example, overload control can be analyzed using the queuing systems theory (see Abaev et al. (2014); Gaidamaka et al. (2014)).

For modeling we used the continuous (fluid) model for TCP and RED (see Demidova et al. (2014); Eferina et al. (2014)). However, this approach allowed us to model the TCP protocol

only partially. In addition, we received differential inclusions instead of differential equations.

We used software package ns-2 for the resulting model verification. Discrete event simulator ns-2 (see Altman and Jiménez (2012); Issariyakul and Hossain (2012)) implements some network protocols. Because of this, it has a low scalability and is suitable for modeling of small networks at small time intervals.

Also the verification on the basis of the software router (see Velieva et al. (2014)) was carried out.

For further development of our model, it was decided to use a hybrid approach (see Maler (1992, 2002); Färnqvist et al. (2002); Hespanha et al. (2001); Bohacek and Lee (2001)). This article discusses a general approach to a hybrid modeling of TCP and RED. The implementation is demonstrated on the basis of Modelica (see Fritzon (2003, 2011)) language with the help of OpenModelica system.

The structure of the article is as follows. The first section describes the hybrid paradigm of mathematical modeling. The general information about the language of physical modeling named Modelica is given in the same section. Modelica implements continuous and hybrid paradigms. The TCP Reno protocol as also continuous and hybrid models of this protocol are presented in the second section. Similarly, in the third section this is done for the RED active traffic management module. The last section describes the network topology and the overall structure of the hybrid model for its implementation.

THE HYBRID APPROACH TO MODELING

The hybrid¹ (see Maler (1992, 2002); Färnqvist et al. (2002); Hespanha et al. (2001); Bohacek and Lee (2001)) system has both continuous and discrete aspects of behavior. The hybrid behavior may be due to different reasons.

- Hybrid behavior is due to the joint operation of the continuous and discrete objects. For example, the automatic control system with continuous control object and discrete control device.
- Hybrid behavior is caused by changes in the structure of the system. A system with variable number of components may be considered as an example.
- Hybrid behavior may be caused by instant qualitative changes in a continuous object. In this case, qualitative changes during the simulation of continuous systems are presented as discrete events. As a result, the hybridism is not an inherent characteristic of the system, but the modeling technique.

Hybrid systems may be considered as discrete-continuous or continuous-discrete systems.

- The waiting time for the next input and the duration of the output action can be taken into account in discrete systems.

¹Other names are *continuous-discrete system*, *system with variable structure*, *event-driven system*.

- The coexistence of instant and long-term processes in a continuous-time model.

We will add discrete elements to the initially developed continuous dynamic model.

Discrete events in continuous dynamic models can be created by the following components:

- initial conditions and parameter values in the right sides;
- the form of the right sides;
- the number of equations.

The change of the initial conditions and the stepwise change of the parameters may be considered as the same type changes because the stepwise change of parameters can be described as a replacement of the initial conditions for a new system of equations.

Thus, one can use both indicator functions and differential inclusions within the hybrid model. This technique allows to replace the system with a variable right side by the system with the constant right side and the variable initial conditions.

For example, let's suppose a system of differential equations with a piecewise constant parameter π :

$$\begin{aligned} \frac{dx}{dt} &= f(x, t, \pi), \\ \pi &= \begin{cases} \psi_1, & x \in \mathfrak{X}_1, \\ \psi_2, & x \in \mathfrak{X}_2. \end{cases} \end{aligned}$$

Then it can be replaced by the following system:

$$\begin{cases} \frac{dx}{dt} = f(x, t, \pi), \\ \frac{d\pi}{dt} = 0, \end{cases}$$

with the following initial conditions:

$$\begin{cases} \pi(0) = \psi_1, & x \in \mathfrak{X}_1, \\ \pi(0) = \psi_2, & x \in \mathfrak{X}_2. \end{cases}$$

This technique may be applied in modeling of the behavior of the TCP protocol and RED mechanism in different states.

Modelica modeling language

Modelica language (see Fritzon (2003, 2011)) is developed by a nonprofit organization Modelica, which also develops a free library for this language. Modelica is positioned as an object-oriented physical modeling language.

The classes are the basis of modeling in Modelica. Modelica class includes not only the fields and methods, but also the equations which relate variables to each other. The equations are a special entity in the language. The number of equations and variables in the program must be the same. Also, the fields may have a different type of variability: a constant parameter (does not change under the current modeling), a variable. The field in the class can be both the object of embedded types and the object of user-defined types. The classes can be inherited, the entire contents of inherited class is copied to the inheriting class, including equations.

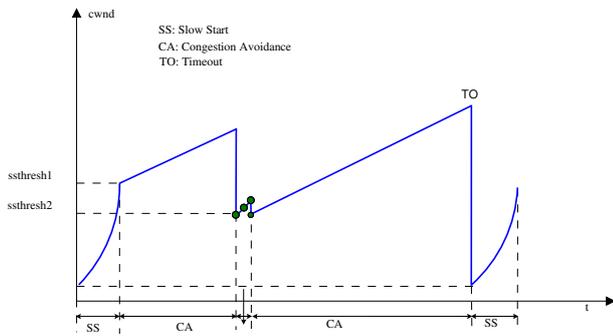


Fig. 1. TCP phases

Modelica supports continuous and hybrid (continuous-discrete) paradigms. However, the discrete elements are also present in the language.

Modelica language is presented by a large number of commercial implementations such as:

- Dymola [<http://www.claytex.com/>];
- CATIA [<http://www.3ds.com/products-services/catia/>];
- MapleSim [<http://www.maplesoft.com/products/maplesim/>];
- Wolfram SystemModeler [<http://www.mathcore.com/>].

There are also open source implementations of Modelica language and the environment:

- OpenModelica [<https://openmodelica.org/>];
- Scicos [<http://www.scicos.org/>];
- JModelica.org [<http://jmodelica.org/>].

HOW TCP WORKS

The TCP protocol uses a sliding window mechanism to avoid a congestion. The implementation of this mechanism depends on the particular type of TCP protocol.

TCP congestion control mechanism

Since the original model (see Misra et al. (1999, 2000); Velieva et al. (2014)) is based on the TCP Reno protocol, then this particular protocol will be simulated.

In TCP Reno protocol the congestion control mechanism consists of the following phases: slow start, congestion avoidance, fast transfer and fast recovery. Dynamics of changes in congestion window size (CWND) depends on the specific phase (see Fig. 1).

Each time the source receives a delivery notification (Acknowledgement, ACK), in slow start phase congestion window is increased. The source increases the congestion window size depending on the number of confirmed segments (Segment Size, SS): $cwnd = cwnd + 1$ for each transmitted ACK. Initial congestion window size (Maximum Segment Size, MSS) can take a value of 1, 2 or 10 segments. The receiver sends an

ACK for each packet, but in reality it can be assumed that confirmation come together at the end of the double turnaround time (Round-Trip Time, RTT). Thus, the congestion window is doubled after the round-trip time.

When TCP Reno window size takes the certain value the protocol mechanism enters the congestion avoidance phase. In this phase, the congestion window is increased by the amount of $1/cwnd$ for each acknowledgment ACK, which is equivalent to increasing of the window by one packet for the double-turn.

Protocol TCP Reno monitors two options of packet loss:

- Triple Duplicate ACK (TD). Let n -th package is not delivered, and subsequent packets ($n + 1$, $n + 2$, etc.) are delivered. For each packet delivered in violation of prioritization (for $n + 1$, $n + 2$, and so on), the recipient sends ACK message for the last undelivered (n -th) package. With receiving three such packets the source resends the n -th package. In addition, the window size is decreased by 2 times $cwnd \rightarrow cwnd/2$.
- Timeout (TO). While sending a package the timeout timer is started. After receiving the confirmation the timer is restarted. Wherein the window size is set to the initial value of the congestion window. The first lost package is resent. The protocol passes into a slow start phase.

Overall congestion control algorithm belongs to AIMD algorithms type (Additive Increase, Multiplicative Decrease) — an additive increase of the window size and multiplicative decrease of it.

The transition to the continuous model for the congestion window

Because we want to construct the hybrid continuous-discrete model, we need pass to the continuous-time model in order to describe the operation of each TCP phase. The transition between the phases will be described by discrete states.

Using the results of section , the behavior of our model may be formalized. A congestion window change is described by an elementary event, which corresponds to a single acknowledgment or confirmation of all. Let us assume that the elementary event is the arrival of all acknowledgments that occurs during the round-trip time (RTT).

In the slow start phase a congestion window size increases with each occurrence of confirmation (ACK):

$$W(t_n^{ACK} + \Delta t^{ACK}) = W(t_n^{ACK}) + 1. \quad (1)$$

We now rewrite (1) relative to the round-trip time T :

$$W(t_n + \Delta t) = W(t_n) + 1 \cdot W(t_n),$$

$$\frac{W(t_n + \Delta t) - W(t_n)}{\Delta t} = \frac{W(t_n)}{\Delta t}.$$

Assuming that $\Delta t = T$, we obtain

$$\frac{dW}{dt} = \frac{W}{T},$$

$$d \ln W = \frac{dt}{T}, \quad \ln W = \frac{1}{T}; \quad W = \exp\left\{\frac{1}{T}\right\}.$$

Thus, the window grows exponentially, as it should be in a slow start phase in accordance with the TCP description.

Similarly, we examine congestion avoidance phase. For each occurrence of an ACK the window size is increased:

$$W(t_n^{ACK} + \Delta t^{ACK}) = W(t_n^{ACK}) + \frac{1}{W(t_n^{ACK})}.$$

We rewrite this for a round-trip time:

$$W(t_n + \Delta t) = W(t_n) + \frac{1}{W(t_n)}W(t_n);$$

$$\frac{W(t_n + \Delta t) - W(t_n)}{\Delta t} = \frac{1}{\Delta t}.$$

Assuming that $\Delta t = T$, we have

$$\frac{dW}{dt} = \frac{1}{T},$$

$$dW = \frac{dt}{T}, \quad W = \frac{1}{T}.$$

The result is a linear increase of the window, as described in the specification of the TCP.

Construction of hybrid model for TCP

To construct a hybrid model we need:

- to write a dynamic model for each state (done in section);
- to replace the system with step parameters by the system with variable initial conditions;
- to write the state diagram (Fig. 2).

The resulting chart (Fig. 2) can be converted to a Modelica program. We give a fragment of the listing. Here we demonstrate only the state transition algorithm for TCP. As can be seen, it is made by almost verbatim copying of UML-diagrams (it is theoretically possible to carry out the code generation based on the corresponding chart). For greater clarity, the certain variables before the algorithm² are given.

RED ADAPTIVE CONGESTION CONTROL MECHANISM

To improve the channel performance the queue management on routers needs to be optimized. One of the possible approaches is to use the random early detection algorithm (Random Early Detection, RED) (see Floyd and Jacobson (1993)) or RED modifications.

Generally, the RED algorithm is very simple, but at the same time it presents the effective model of active traffic management. In addition, it imposes virtually no restrictions on the researcher. As a result, researchers are constantly creating new modifications of this algorithm (see Floyd and Jacobson (1993); Feng et al. (2015); Lautenschlaeger and Francini (2015); Karmeshu et al. (2016)).

Listing 1. State transition algorithm for TCP protocol

```

model TCPSEnder "Transmission Control
  Protocol"
  // Variables
  type TCPState = enumeration(slowStart,
    fastRecov, congestAvoid, timeOut);
  discrete TCPState state(start =
    TCPState_slowStart);
  parameter Real timeout_th = 4.0 "
    Window threshold for entering
    timeout";
  Real ssth(start = 32.0) "Slow start
    thresholds";
  Real w(start = MinSS, min = 1, max =
    w_max) "Communication window sizes,
    no of packets";
  Real drop_timer(start = 0) "Drop timer
    ";
  Real retr_timer(start = 0) "
    Retransmission timer";
  Boolean DelayD(start = false) " Delay
    induced Drop detected";

  //
  // Some code
  //
algorithm
  // State transitions
  state := TCPState_slowStart;
  when edge(DelayD) and w >= timeout_th
    and (state == TCPState_slowStart or
    state == TCPState_congestAvoid)
    then
    state := TCPState_fastRecov;
  elsewhen w >= ssth and state ==
    TCPState_slowStart then
    state := TCPState_congestAvoid;
  elsewhen w < timeout_th and edge(
    DelayD) and (state ==
    TCPState_slowStart or state ==
    TCPState_congestAvoid) then
    state := TCPState_timeOut;
  elsewhen retr_timer < 0 and state ==
    TCPState_fastRecov then
    state := TCPState_congestAvoid;
  elsewhen retr_timer < 0 and state ==
    TCPState_timeOut then
    state := TCPState_slowStart;
  end when;
end TCPSEnder;

```

²Note that the function *edge()* (used only with *Boolean* variables) is set to *true* only in case the operand only just received value *true* (that is, its previous value was *false*).

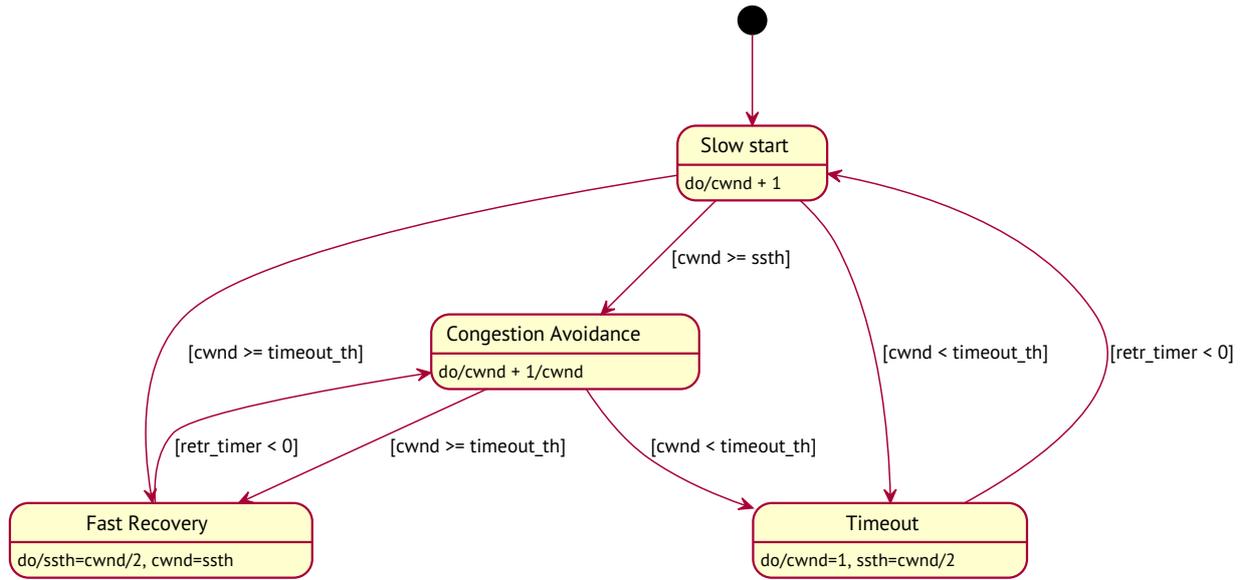


Fig. 2. TCP state diagram

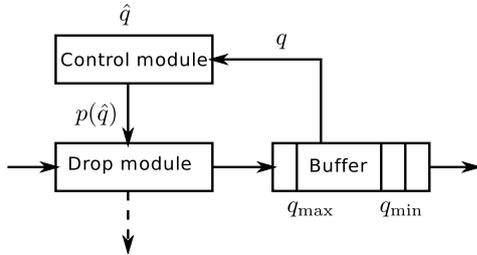


Fig. 3. RED module

The previous implementation of the RED algorithm (see Velieva et al. (2014)) was done in the ideology of continuous modeling. Therefore, we were limited to the following model assumptions:

- we have considered only the congestion avoidance stage;
- we took into account the loss of packets only by triple duplication.

The hybrid modeling allows us to overcome this barrier and to investigate the mechanism in its entirety.

How RED works

The module that implements the RED-type algorithm can be schematically represented as follows (fig. 3):

RED algorithm uses a weighted value of the queue length \hat{Q} as factor for the determination of packet dropping probability. In order to calculate \hat{Q} the exponentially weighted moving-average (EWMA) is used:

$$\hat{Q}_{k+1} = (1 - w_q)\hat{Q}_k + w_q Q_k, \quad k = 0, 1, 2, \dots,$$

where w_q , $0 < w_q < 1$ is a weight coefficient of the exponentially weighted moving-average.

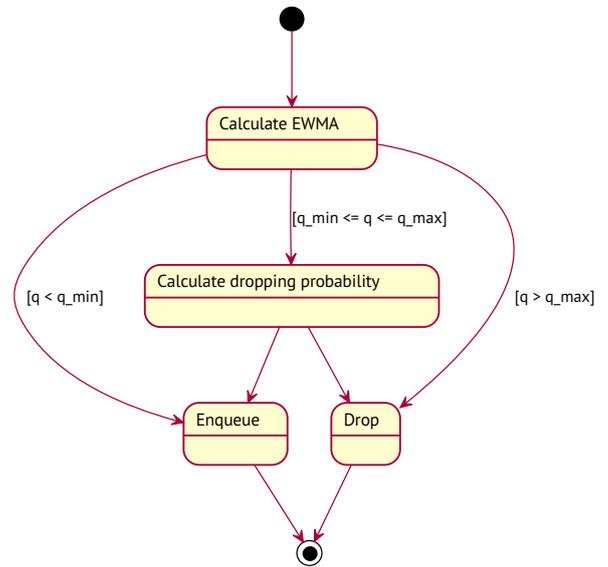


Fig. 4. RED state diagram

As the the average queue length increases, the packets drop probability also increases (see. (2)). For dropping management algorithm the dropping function with two average queue length thresholds Q_{min} and Q_{max} as arguments (Fig. 5) is used:

$$p(\hat{Q}) = \begin{cases} 0, & 0 < \hat{Q} \leq Q_{min}, \\ \frac{\hat{Q} - Q_{min}}{Q_{max} - Q_{min}} p_{max}, & Q_{min} < \hat{Q} \leq Q_{max}, \\ 1, & \hat{Q} > Q_{max}. \end{cases} \quad (2)$$

Here $p(\hat{Q})$ is the package dropping function, \hat{Q} is the queue length weighted average, Q_{min} and Q_{max} are thresholds of queue length weighted average, p_{max} is the maximum level of packages reset.

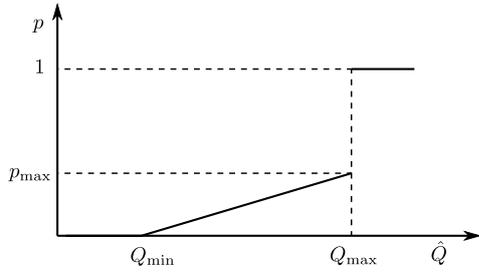


Fig. 5. RED drop function

The transition to the continuous model for RED algorithm

The behavior of an exponentially weighted moving average queue length, which is a constraint equation between the source and the receiver, is described further.

Let w_q is the parameter of the moving average (see Floyd and Jacobson (1993)). Based on the formula for exponentially weighted moving average, we can write:

$$\begin{aligned}\hat{Q}(t_n + \Delta t) &= (1 - w_q)\hat{Q}(t_n) + w_q Q(t_n), \\ \hat{Q}(t_n + \Delta t) - \hat{Q}(t_n) &= -w_q \hat{Q}(t_n) + w_q Q(t_n), \\ \frac{\hat{Q}(t_n + \Delta t) - \hat{Q}(t_n)}{\Delta t} &= \frac{w_q}{\Delta t} (Q(t_n) - \hat{Q}(t_n)).\end{aligned}$$

Let $\delta = \Delta t$. Let us give δ meaning of time spent by one packet in the queue. Assuming that C is the service intensity, we can write $\delta = \frac{1}{C}$. Then the equation for the exponentially weighted moving average queue length takes the form:

$$\frac{d\hat{Q}}{dt} = \frac{w_q}{\delta} (Q - \hat{Q}) = w_q C (Q - \hat{Q}).$$

Construction of hybrid model for RED algorithm

To construct a hybrid model, we need:

- to write a dynamic model for each state;
- to replace the system with step parameters by the system with variable initial conditions;
- to write the state diagram (Fig. 4).

In general RED model is more simple than TCP model. Especially considering that the RED model has only one state, hence there is no need to encode the transitions between states.

THE GENERAL SCHEME OF THE MODEL

For the study of interaction between active queue management module and the TCP traffic source we will use a simple dumbbell topology (Fig. 6). For all its simplicity, it gives a good description of the basic elements of our model.

In Modelica language the model is written as follows (see listing 2). Routers play only the role of connectors, without any other functionality. RED module is configured as a separate class. To set a delay in the network, we introduced the links.

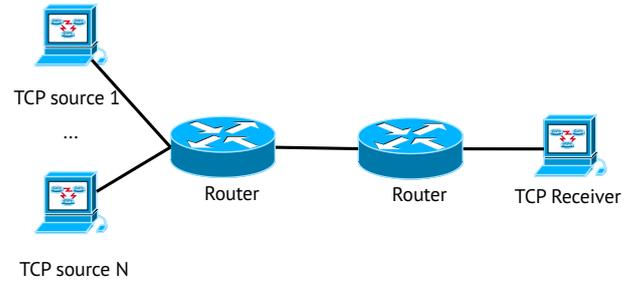


Fig. 6. Dumbbell topology

Listing 2. Implementation of the topology

```
model RED_hybrid "RED hybrid model"
  Router router1;
  Router router2;
  TCPSender aTCP(L = 500);
  Receiver receiver;
  REDqueue red_queue;
  Link link1(delay = 0.02);
  Link link2(delay = 0.02);
equation
  connect(aTCP.o, link1.i);
  connect(link1.o, router1.i);
  connect(router1.o, red_queue.i);
  connect(red_queue.o, router2.i);
  connect(router2.o, link2.i);
  connect(link2.o, receiver.i);
end RED_hybrid;
```

CONCLUSIONS

We investigated the RED active traffic management mechanism as the implementation of the system with a threshold control. Some mathematical models (both analytical and simulation) of this mechanism using different paradigms and techniques were presented. On closer inspection, the presented modeling techniques have shown their shortcomings.

The considered in the article hybrid (continuous-discrete) approach seems to be the most appropriate for network protocol modeling.

A hybrid approach can be used both in the analytical modeling, and in simulations. A hybrid approach can be used both in the analytical modeling, and in simulations. Unfortunately, this approach does not actively used by researchers, although it is implemented in a number of computer simulation systems.

In the future, it seems useful to develop a library for simulation of the interaction between different types of TCP protocol with different versions of RED algorithm.

a) Notes and Comments: The work is partially supported by RFBR grants No's 14-01-00628, 15-07-08795, and 16-07-00556.

REFERENCES

- Abaev, P., Gaidamaka, Y., Samouylov, K., Pechinkin, A., Razumchik, R. and Shorgin, S. (2014), Hysteretic control technique for over-

- load problem solution in network of SIP servers, *Computing and Informatics* 33(1), 218–236.
- Altman, E. and Jiménez, T. (2012), NS Simulator for Beginners, *Synthesis Lectures on Communication Networks* 5(1), 1–184.
- Bohacek, S. and Lee, J. (2001), Analysis of a TCP hybrid model, Proc. of the 39th Annual Allerton Conference on Communication, Control, and Computing, pp. 1–10.
- Demidova, A. V., Korolkova, A. V., Kulyabov, D. S. and Sevastyanov, L. A. (2014), The method of constructing models of peer to peer protocols, 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, pp. 557–562.
- Eferina, E. G., Korolkova, A. V., Gevorkyan, M. N., Kulyabov, D. S. and Sevastyanov, L. A. (2014), One-Step Stochastic Processes Simulation Software Package, *Bulletin of Peoples Friendship University of Russia. Series “Mathematics. Information Sciences. Physics”* (3), 46–59.
- Färnqvist, D., Strandemar, K., Johansson, K. H. and Hespanha, J. P. (2002), Hybrid Modeling of Communication Networks Using Modelica, The 2nd International Modelica Conference, pp. 209–213.
- Feng, C.-W., Huang, L.-F., Xu, C. and Chang, Y.-C. (2015), Congestion Control Scheme Performance Analysis Based on Nonlinear RED, *IEEE Systems Journal* pp. 1–8.
- Floyd, S. and Jacobson, V. (1993), Random Early Detection Gateways for Congestion Avoidance, *IEEE/ACM Transactions on Networking* 1(4), 397–413.
- Fritzson, P. (2003), *Principles of Object-Oriented Modeling and Simulation with Modelica 2.1*, Wiley-IEEE Press.
- Fritzson, P. (2011), *Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica*, John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Gaidamaka, Y., Pechinkin, A., Razumchik, R., Samouylov, K. and Sopin, E. (2014), Analysis of an M—G—1—R queue with batch arrivals and two hysteretic overload control policies, *International Journal of Applied Mathematics and Computer Science* 24(3), 519–534.
- Hespanha, J. P., Bohacek, S., Obraczka, K. and Lee, J. (2001), Hybrid Modeling of TCP Congestion Control, *Lncs*, number 2034, pp. 291–304.
- Issariyakul, T. and Hossain, E. (2012), *Introduction to network simulator NS2*, Vol. 9781461414.
- Karmeshu, Patel, S. and Bhatnagar, S. (2016), Adaptive Mean Queue Size and Its Rate of Change: Queue Management with Random Dropping, pp. 1–17.
- Lautenschlaeger, W. and Francini, A. (2015), Global Synchronization Protection for Bandwidth Sharing TCP Flows in High-Speed Links, Proc. 16-th International Conference on High Performance Switching and Routing, IEEE HPSR 2015, Budapest, Hungary.
- Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994), On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Transactions on Networking* 2(1), 1–15.
- Maler, O. (1992), Hybrid Systems and Real-World Computations, Workshop on Theory of Hybrid Systems, Springer-Verlag, Lyndby, Denmark.
- Maler, O. (2002), Control from computer science, *Annual Reviews in Control* 26(2), 175–187.
- Misra, V., Gong, W.-B. and Towsley, D. (1999), Stochastic differential equation modeling and analysis of TCP-window size behavior, *Proceedings of PERFORMANCE 99*.
- Misra, V., Gong, W.-B. and Towsley, D. (2000), Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED, *ACM SIGCOMM Computer Communication Review* 30(4), 151–160.
- Paxson, V. and Floyd, S. (1995), Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking* 3(3), 226–244.
- Paxson, V. and Floyd, S. (1997), Why we don’t know how to simulate the Internet, Proceedings of the 29th conference on Winter simulation - WSC ’97, ACM Press, New York, New York, USA, pp. 1037–1044.
- Velieva, T. R., Korolkova, A. V. and Kulyabov, D. S. (2014), Designing installations for verification of the model of active queue management discipline RED in the GNS3, 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, pp. 570–577.

AUTHOR BIOGRAPHIES

ANNA V. KOROLKOVA received his Ph.D. in Mathematics in 2010. Since then, she has worked as associate professor in Peoples’ Friendship University of Russia. Her current research activity focuses on mathematical modeling. Her email address is akorolkova@sci.pfu.edu.ru.

DMITRY S. KULYABOV received his Ph.D. in Physics in 2000. Since then, he has worked as associate professor in Peoples’ Friendship University of Russia. His current research activity focuses on mathematical modeling. His email address is yamadharma@gmail.com.

LEONID A. SEVASTIANOV received his D.Sc. in Phys.-Math. in 1999. Since then, he has worked as full professor in Peoples’ Friendship University of Russia. His current research activity focuses on mathematical modeling. His email address is leonid.sevast@gmail.com.

TATYANA R. VELIEVA postgraduate student in Peoples’ Friendship University of Russia. Her current research activity focuses on mathematical modeling. Her email address is trvelieva@gmail.com.

PAVEL O. ABAEV received his Ph.D. in Computer Science from the Peoples’ Friendship University of Russia in 2011. He is an Assistant Professor in the Department of Applied Probability and Informatics at Peoples’ Friendship University of Russia since 2013. His current research focus is on Software-Defined Network, performance analysis of wireless 5G networks and M2M communications, applied probability and queuing theory, and mathematical modeling of communication networks. His email address is pabaev@sci.pfu.edu.ru.

SIR ANALYSIS IN SQUARE-SHAPED INDOOR PREMISES

A. Samuylov[†], D. Moltchanov[†], Yu. Gaidamaka*, V. Begishev*, R. Kovalchukov*, P. Abaev*, S. Shorgin[‡]

*Peoples' Friendship University of Russia, Moscow, Russia,

Email: {ygaidamaka, vobegishev, rnkvalchukov, pabaev}@sci.pfu.edu.ru

[†]Tampere University of Technology, Tampere, Finland,

Email: {andrey.samuylov, dmitri.moltchanov}@tut.fi

[‡]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia,

Email: sshorgin@ipiran.ru

KEYWORDS

Wireless networks, signal-to-interference ratio, distributions, indoor propagation, square cells, wall penetration

ABSTRACT

The increased wireless network densification has resulted in availability of wireless access points (AP) in almost each and every indoor location (room, office, etc.). To provide complete in-building coverage very often an AP is deployed per room. In this paper we analyze signal-to-interference (SIR) ratio for wireless systems operating in neighboring rooms separated by walls of different materials by explicitly taking into account the propagation and wall penetration losses. Both AP and direct device-to-device (D2D) configurations are addressed. Our numerical results indicate that the performance of such system is characterized by both the loss exponent describing the propagation environment of interest and wall materials. We provide the numerical results for typical wall widths/materials and analyze them in detail.

INTRODUCTION

The predicted increase in the user traffic demands places extreme requirements on the future evaluation of mobile systems, often referred to as fifth generation (5G) networks [1], [2]. In addition to physical layer improvements including advanced modulation and coding and MIMO techniques, over the last decade researchers investigated a number of network solutions providing decisive performance improvements including the use of small (micro/pico/femto) cells [3], client-relays [4], direct in-band and out-of-band device-to-device communications [5]. All these concepts target aggressive spatial reuse of frequencies promising substantial area capacity gains.

With the adoption of novel mechanism the user devices are expected to take a more active part in 5G systems and, in some cases, even take on the role of the network infrastructure in providing wireless connectivity such as offering D2D-based data relaying, proximity services, etc. This shift from the classic cellular model is dictated by the progress in communications technologies: the user devices are augmenting their capabilities, whereas the base stations (BSs) are becoming smaller as a result of the ongoing network densification [6].

The networks densification, novel networking and service mechanisms as well as the trend to use multiple access technologies to serve the users, known as heterogeneous cellular system concept, altogether lead to increased randomness of the network, where the positions of servicing stations such as BSs, relays and D2D partners are random rather than deterministic.

The signal-to-interference ratio (SIR) is a universal metric specifying performance of wireless systems [7]. Once SIR is known one could describe the Shannon rate of the channel and spectral efficiency of the system. In contrast to noise-limited systems, where the bit error rate (BER) decreases exponentially with signal-to-noise ratio (SNR), the heterogeneous mobile networks are interference-limited showing linear improvement of BER with respect to SIR. Thus, the increase of the emitted power does not improve the performance of these systems. Thus, the problem of finding SIR for typical network configurations is of special importance characterizing applicability and typical scenarios of modern and future wireless technologies.

The SIR performance of wireless systems is often studied using the tools of stochastic geometry [8]. The basic approach is to specify the point process on the plane modeling positions of the stations and then derive the interference at the point of interest. The last step is rather complex as we need closed-form distribution of distance to the point of interest from at least several neighboring points. For this reason typical considered models are often limited to Poisson point process on the plane for which we immediately have closed-form expressions for distributions of distances to the i -th neighbor [9].

The constantly increasing need for wireless connectivity on-the-go [10] are gradually changing the way service is provisioned in wireless networks. Nowadays, one of the trends is to deploy small wireless stations including both IEEE 802.11 or micro-LTE access points (AP) in crowded areas to benefit from increased network densification [6] and shorter propagation distances. Examples include large shopping mall, office environment, where one of few adjacent rooms is served by an AP having relatively small coverage area. In this dense environment interference between neighboring APs is inevitable and may easily lead to degraded system performance.

In this paper, using the tools of stochastic geometry, we analyze performance of wireless systems operating in neighboring rooms of rectangular configuration. We consider both direct device-to-device and AP configurations assuming that

the systems in adjacent rooms operate at the same frequency. The analytical results are compared to simulations showing adequate agreement. Numerical results for the set of input metrics demonstrate that the system performance is dictated by the interplay between path loss exponent typical for a given environment and type of the walls used between rooms.

The rest of the paper is organized as follows. First, in the next section we introduce a system model. Further, we analytically study SIR for downlink scenario. The simulation models for both downlink and D2D scenarios are introduced next. Numerical results for different sets in input variables are illustrated. Conclusions are drawn in the last section.

SYSTEM MODEL

In this study we focus on an indoor scenario with grid aligned rooms, see Fig. 1 that are typical for shopping malls or office buildings. In these environments rooms are often of rectangular or square shapes. Each room is assumed to be equipped with an AP deployed in the geometrical center. To take advantage of the wireless network densification trend as a solution to upgrade the degree of spatial reuse, the devices in adjacent rooms are assigned the same set of communication channels [6]. The mobile terminals (users) operating over the same channel are assumed to be uniformly distributed over the room, one per room. We concentrate on the so-called tagged user in the central room, see Fig. 1(a) and Fig. 1(b). We assume both AP and users to be equipped with omnidirectional antennas. We do not focus on a particular radio technology addressing the general case.

In addition to AP scenario we also address D2D configuration, sketched in Fig. 1(c). The principal difference compared to AP case is that both transmitter and receiver are assumed to be uniformly distributed within a room. Under this assumption the configuration is symmetric, i.e., we do not have to distinguish between uplink and downlink cases. Similarly, we concentrate on D2D pair located in the central room.

Focusing on SIR, as a metric of interest, for both AP and D2D configurations we calculate it for a randomly chosen receiving device, taking into account the interference from a set of neighboring rooms. Using the commonly used propagation model, we add a correction factor, accounting for the attenuation of a signal when passing through a wall

$$SIR = \frac{S}{\sum_{i=1}^N (I_i B_i)}, \quad (1)$$

where S is the received signal power, N is the number of interfering sources, I_i is the interference power from the i^{th} source, B_i is the correction factor.

The received signal power is a function of the distance between the transmitter and the receiver (or between the interfering device and the device of interest). The functions in (1) are specified as

$$S = S(l) = gl^{-\gamma}, \quad I_i = I_i(l_i) = gl_i^{-\gamma}, \quad (2)$$

where g is the transmit power assumed to be constant for all the transmitters, l is the distance, and γ is the path loss exponent, which ranges from 2 to 6.

ANALYTICAL APPROACH

Uplink scenario

Consider the case with four interfering devices. Here we build an analytical model for Uplink scenario where devices are located in the square rooms with sides of $c = a_j = b_j$, $j = \overline{1, 3}$, as shown in Fig. 1(a). Taking this into account, (1) can be simplified as

$$SIR = \frac{S(R_0)}{\sum_{i=1}^4 (I_i B_i)}, \quad (3)$$

$$S(R_0) = gR_0^{-\gamma}, \quad \sum_{i=1}^4 I_i(D_i B_i) = g \sum_{i=1}^4 D_i^{-\gamma} B_i, \quad (4)$$

where R_0 is the distance between Tx_0 and Rx_0 . The distance between Tx_i and Rx_0 is denoted by D_i . Assuming constant transmit power (3) reads as

$$SIR = \frac{gR_0^{-\gamma}}{g \sum_{i=1}^4 (D_i^{-\gamma} B_i)} = \frac{R_0^{-\gamma}}{\sum_{i=1}^4 (D_i^{-\gamma} B_i)}. \quad (5)$$

Introduce the following random variables

$$\xi = SIR, \quad \eta_1 = R_0^{-\gamma}, \quad \eta_2 = D^{-\gamma}. \quad (6)$$

According to the method described in [11] the probability density function (pdf) of $D^{-\gamma}$, $W_{\eta_2}(y_2)$, is given by

$$W_{\eta_2}(y_2) = \left(\frac{2}{\gamma c^2} y_2^{\frac{-2}{\gamma}-1} \right) \times \begin{cases} 0, & y_2 \geq \left(\frac{c}{2}\right)^{-\gamma} \\ \arcsin \left[\frac{c}{2y_2^{-1/\gamma}} \right] - \arcsin \left[\frac{\sqrt{-9c^1 + 4y_2^{-2/\gamma}}}{2y_2^{-1/\gamma}} \right], & \left(c\sqrt{\frac{5}{2}}\right)^{-\gamma} < y_2 \leq \left(\frac{3c}{2}\right)^{-\gamma} \\ \arcsin \left[\frac{\sqrt{-c^2 + 4y_2^{-2/\gamma}}}{2y_2^{-1/\gamma}} \right], & \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} < y_2 \leq \left(\frac{c}{2}\right)^{-\gamma} \\ \arcsin \left[\frac{c}{2y_2^{-1/\gamma}} \right], & \left(\frac{3c}{2}\right)^{-\gamma} < y_2 \leq \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} \end{cases} \quad (7)$$

Consider now the total sum of all interfering signals. Since the convolution of (7) is not trivial, we approximate this sum with a Normal distribution. Using (7) we calculate the mean and variance of the interfering signal power as

$$\tilde{\mu} = \sum_{i=1}^4 \mu_i B_i, \quad \tilde{\sigma}^2 = \sum_{i=1}^4 \sigma_i^2 B_i. \quad (8)$$

Since all four rooms are symmetric, the distributions of the individual interfering signals are the same. Assuming the same material and width for all walls between the room of interest and the interfering rooms, (8) is reduced to

$$\tilde{\mu} = 4\mu B, \quad \tilde{\sigma}^2 = 4\sigma^2 B. \quad (9)$$

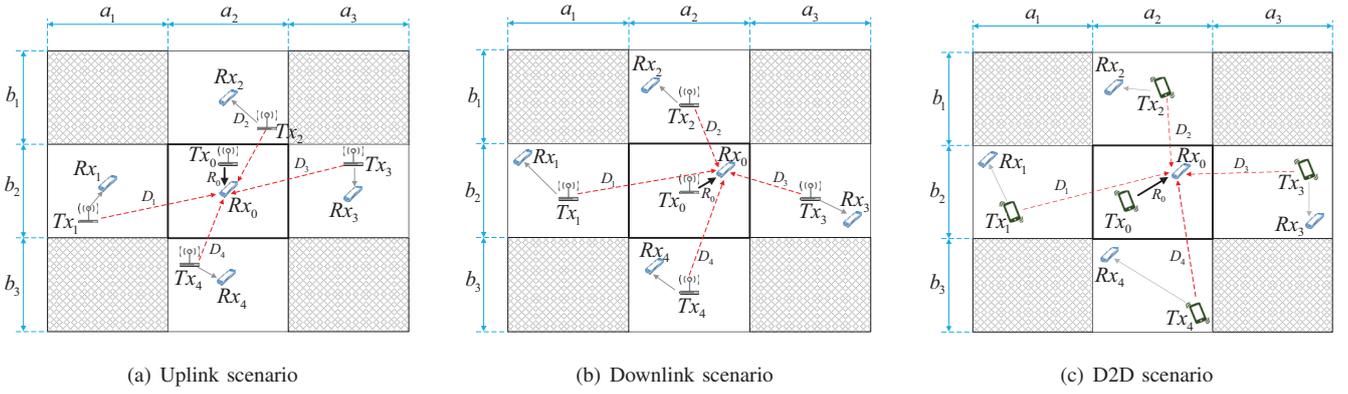


Fig. 1. Layout of the three considered scenarios.

The approximating Normal distribution is then written as

$$N_{\eta_2}(x) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} e^{-\frac{(x-\tilde{\mu})^2}{2\tilde{\sigma}^2}}. \quad (10)$$

The pdf of the signal of interest $W_{\eta_1}(y_1)$ is obtained as

$$W_{\eta_1}(y_1) = \left(\frac{2}{\gamma c^2} y_1^{\frac{-2}{\gamma}-1}\right) \times \begin{cases} \pi, & \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} < y_1 < \infty \\ 2 \left(\arcsin \left[\frac{c}{2y_1^{-1/\gamma}} \right] - \arcsin \left[\frac{\sqrt{-c^2 + 4y_1^{-2/\gamma}}}{2y_1^{-1/\gamma}} \right] \right), & \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} < y_1 \leq \left(\frac{c}{2}\right)^{-\gamma} \\ \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} < y_1 \leq \left(\frac{c}{2}\right)^{-\gamma}. \end{cases} \quad (11)$$

Now we derive the density function for (3) by applying the functional transformation of random variables [12]. We obtain the pdf of random variable ξ denoting the SIR based on the joint pdf of the signal power of interest and total interfering signal power. Recalling the form of SIR in (3), we establish

$$\begin{aligned} y_3 &= f(x_3, x_4) = x_3/x_4 \\ y_4 &= x_4, \end{aligned} \quad (12)$$

where y_4 is an auxiliary variable.

The inverse of (12) takes the following form

$$\begin{aligned} x_3 &= \varphi(y_3, y_4) = y_3 y_4, \\ \frac{\partial \varphi(y_3, y_4)}{\partial y_3} &= y_4. \end{aligned} \quad (13)$$

The pdf we are looking for is obtained using

$$\begin{aligned} W_{\eta_1, \eta_2}(y_3, y_4) &= \\ &= w_{\chi_1, \chi_2}(\varphi(y_3, y_4), y_4) \left| \frac{\partial \varphi(y_3, y_4)}{\partial y_3} \right|, \end{aligned} \quad (14)$$

where $w_{\chi_1, \chi_2}(y_3, y_4)$ is the joint density of $D^{-\gamma}$ and $\sum R_0^{-\gamma}$. To obtain a univariate density function of SIR we integrate out y_4 in (14)

$$\begin{aligned} W_{\xi}(y_3) &= \\ &= \int_{\mathbf{Y}} w_{\chi_1, \chi_2}(\varphi(y_3, y_4), y_4) \cdot \left| \frac{\partial \varphi(y_3, y_4)}{\partial y_3} \right| dy_4, \end{aligned} \quad (15)$$

where \mathbf{Y} is the range of the variable y_4 for the inverse.

According to the limits imposed by (7, 10), \mathbf{Y} is

$$\begin{aligned} &\left\{ \left(\frac{c}{\sqrt{2}}\right)^{-\gamma} < y_1 \leq \left(\frac{c}{2}\right)^{-\gamma}, -\infty < y_2 < \infty \right\} \cup \\ &\cup \left\{ \left(\frac{c}{2}\right)^{-\gamma} < y_1 < \infty, -\infty < y_2 < \infty \right\}. \end{aligned} \quad (16)$$

Thus, we have

$$\mathbf{Y} = \mathbf{Y}^1 \cup \mathbf{Y}^2, \quad (17)$$

where

$$\begin{aligned} \mathbf{Y}^1 &= \left\{ y_3 < 0, \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} < y_4 < \frac{e^{\frac{1}{2}\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \cup \\ &\cup \left\{ y_3 > 0, \frac{e^{\frac{1}{2}\gamma \ln 2 - \gamma \ln c}}{y_3} < y_4 < \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\}. \end{aligned} \quad (18)$$

and

$$\begin{aligned} \mathbf{Y}^2 &= \left\{ y_3 < 0, y_4 < \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \cup \\ &\cup \left\{ y_3 > 0, y_4 > \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\}. \end{aligned} \quad (19)$$

The density of SIR, $W_{\xi}(y_3)$, is now provided by

$$W_{\xi}(y_3) = \begin{cases} \int_{M_1} I_1(y_3, y_4) dy_4 + \int_{M_3} I_1(y_3, y_4) dy_4, & y_3 < 0 \\ \int_{M_2} I_2(y_3, y_4) dy_4 + \int_{M_4} I_2(y_3, y_4) dy_4, & y_3 \geq 0. \end{cases} \quad (20)$$

where the limits of integration are

$$\begin{aligned} M_1 &= \left\{ (y_3, y_4) : \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} < y_4 < \frac{e^{\frac{1}{2}\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \\ M_2 &= \left\{ (y_3, y_4) : \frac{e^{\frac{1}{2}\gamma \ln 2 - \gamma \ln c}}{y_3} < y_4 < \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \\ M_3 &= \left\{ (y_3, y_4) : y_4 < \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \\ M_4 &= \left\{ (y_3, y_4) : y_4 > \frac{e^{\gamma \ln 2 - \gamma \ln c}}{y_3} \right\} \end{aligned} \quad (21)$$

while the integrands have the following forms

$$I_1(y_3, y_4) = \frac{\arcsin \left[\frac{c}{2(y_3 \cdot y_4)^{\frac{-1}{\gamma}}} \right] - \arcsin \left[\frac{\sqrt{-c^2 + 4(y_3 \cdot y_4)^{\frac{-2}{\gamma}}}}{2(y_3 \cdot y_4)^{\frac{-1}{\gamma}}} \right]}{\frac{4y_4}{\gamma c^2} (y_3 \cdot y_4)^{\frac{-2}{\gamma} - 1} \tilde{\sigma} \sqrt{2\pi}} e^{-\frac{(y_4 - \tilde{\mu})^2}{2\tilde{\sigma}^2}},$$

$$I_2(y_3, y_4) = \frac{2\sqrt{2\pi^3} y_4 \cdot \tilde{\sigma} (y_3 \cdot y_4)^{\frac{-2}{\gamma} - 1}}{\gamma c^2 e^{-\frac{(y_4 - \tilde{\mu})^2}{2\tilde{\sigma}^2}}}. \quad (22)$$

As we show in the numerical results section, this yields a pretty loose approximation due to the fact that we take into account the interference only from four sources, which is not large enough to comply with the central limit theorem. Thus, we compare the obtained results with another approximation based on the hyperexponential distribution. In this case we use the truncated two phase hyperexponential distribution to approximate the interference from a single source (7). The total interference power is obtained by convolving the obtained approximate distribution four times. We will omit the derivation, as the method we use to obtain these results is exactly the same like described in this section.

Downlink and D2D cases

The performed analysis for uplink scenario is feasible due to the independence of propagation paths between interferers and receiver. In downlink scenario these distances are no longer independent as all interference paths share the same fixed point - the receiver of interest. Because of this fact, the approximation for the denominator in (1) cannot be used without any further assumptions, thus complicating further derivations. Due to symmetric configuration of the scenario, one may observe that approximation obtained by assuming independence between these distances will result in significant approximation error. The situation is similar for D2D scenario, where the interference paths are again dependent. To facilitate derivation of the SIR densities in these cases we developed simulation environment described in the next section.

SIMULATION ENVIRONMENT

The developed simulation model based on the Monte Carlo method is a software tool implemented in C++ to gather statistics of SIR for a receiver in the central cluster by directly modeling all the underlying random variables. The tool relies on the same propagation model we used in analytical derivations, allowing not only to acquire SIR for downlink and D2D cases, but as well to assess the accuracy of analytical model.

The basic principle of the tool is as follows. We randomly choose all coordinates of devices that interfere and transmit signal to the target device located in the central cluster. Using SIR expression provided in (5) we collect statistics by repeating this process sufficiently many times. Once the statistics is obtained empirical pdf is constructed. Due to insignificant resource usage for all input values of interest we are able to obtain a number of samples that is significantly higher than recommended $n = 1 + \lceil \log_2 N \rceil$. All the characteristics

of random variables have been obtained using conventional statistical methods.

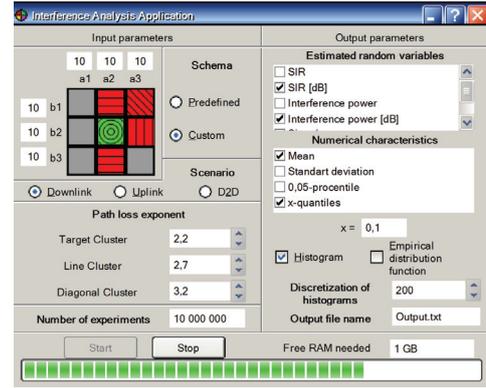


Fig. 2. Simulation tool graphical user interface.

The graphical user interface for Windows operating system is shown in Fig. 2. In the left pane user defines input parameters for modeling. They are the lengths of the sides of rectangular rooms and the value for the path loss exponent. An estimate of the SIR is obtained in either the so-called "standard" mode, when interference from all eight neighboring nodes is taken into account, or in the so-called "custom" mode, where the user manually selects the set of interfering rooms. Fig. 2 demonstrates an example of a user-mode selection.

The simulation model provides three options for the location of the devices in the clusters: the so-called "uplink" scenario, where the transmitters are located in the centers of rooms, and the coordinates of the receivers all follow uniform distribution, the "downlink" scenario, where the receivers are in the centers of their areas and coordinates of the transmitters follow uniform distribution, and the so-called "D2D" scenario, where the coordinates of both receivers and transmitters are distributed uniformly. In the right pane, a user sets the output parameters of the simulator. One or more characteristics can be chosen, including SIR, the power of the useful signal, the power of interfering signal, as well as their statistical characteristics. Particularly, when assessing SIR the 0.05-quantile is of special interest. The representation of the results can be selected by checking the box option "histogram" or "empirical distribution function". The last two parameters are responsible for setting the number of intervals of the histogram and the empirical pdf of random variables.

NUMERICAL RESULTS

In this section we present numerical results for all three considered scenarios. For uplink case we compare analytical data with those obtained using simulations. For comparison purposes, the numerical modeling was carried out in square clusters with sizes $c = 10$ (hereinafter the lengths are given in arbitrary units, e.g. meters), for all three scenarios. For downlink and D2D scenarios simulation data are presented. The values of losses caused by wall penetration are taken from [13]

Fig. 3 illustrates the comparison between analytical and simulation results for room size set to 10 and different path loss exponents. The number of samples used to construct

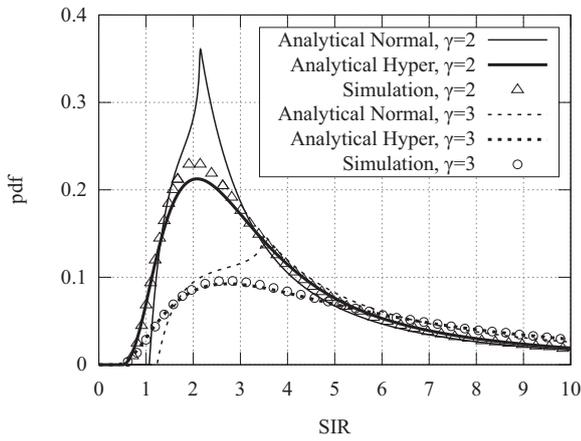


Fig. 3. Comparison of analytical and simulation results.

the empirical density function has been set to $10e6$. First, as one may notice, analytical results obtained by modeling the joint interference by Normal distribution deviate from the simulations implying that the model provides fairly loose approximation. The deviations is attributed to the approximation by the Normal distribution. Basically, it just shifts the distribution of the signal of interest (11), leaving the form intact. This is especially true for small values of SIR that are of special importance in practice specifying the so-called outage probabilities. However, the results obtained using Hyperexponential approximation for a single interferer yield a pretty tight fit for the empirical results and can be used as a lower bound estimate.

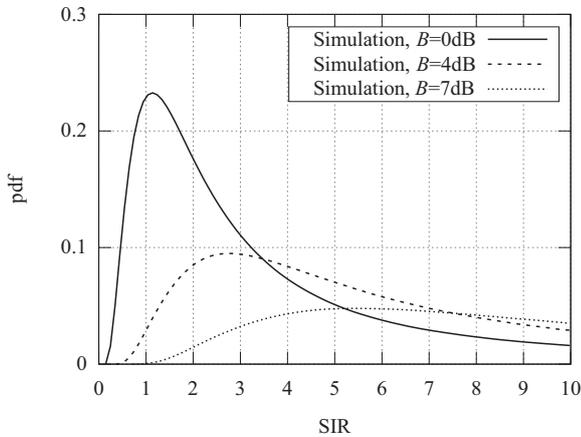


Fig. 4. SIR for uplink scenario, $c = 10$, $\gamma = 3$.

The comparison of SIR densities obtained using the simulation approach for $c = 10$ and path loss exponent $\gamma = 3$, is shown in Fig. 4. These figures were constructed assuming (i) no walls between rooms (loss of 0 dB, i.e. $B = 1$), (ii) wood walls with thickness 102mm resulting in (loss of 4 dB, i.e. $B \approx 0.4$), and (iii) brick walls of thickness 267mm corresponding to (loss of 4 dB, i.e. $B \approx 0.2$). As one may observe, the presence of walls between rooms fundamentally changes the structure of SIR density. As expected, the mode of the distribution shifts to the left implying better system performance.

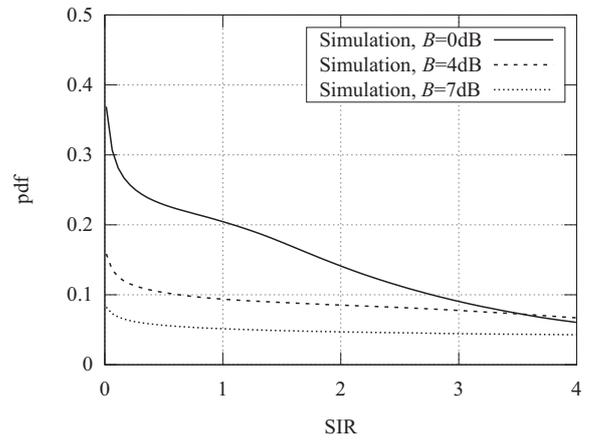


Fig. 5. SIR for downlink scenario, $c = 10$, $\gamma = 3$.

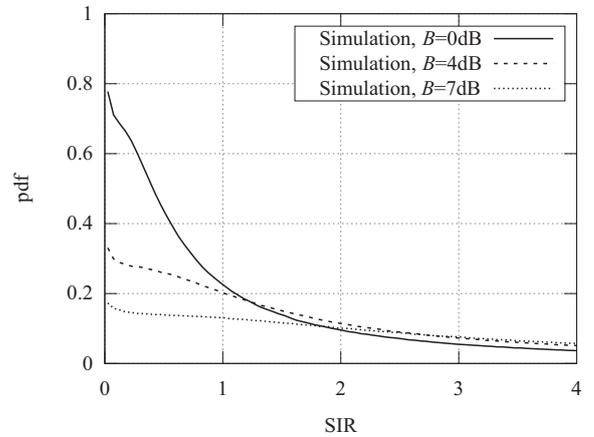


Fig. 6. SIR for D2D scenario, $c = 10$, $\gamma = 3$.

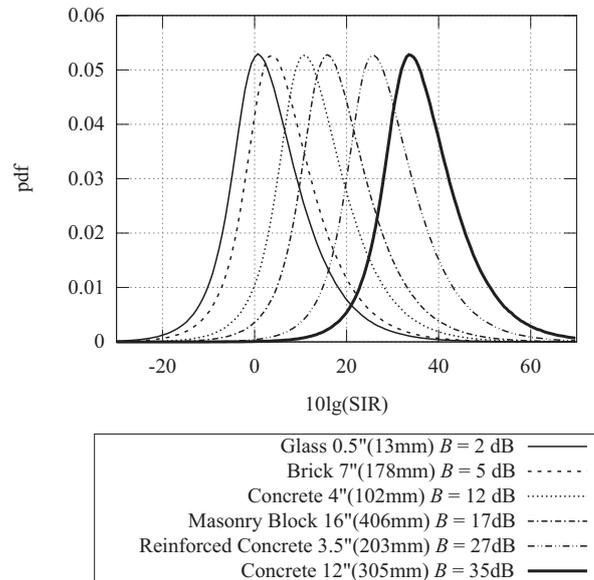


Fig. 7. SIR for a number of different wall materials.

For both downlink and D2D scenarios we obtained SIR densities using simulation approach. Fig. 5 compares SIR distributions for different path loss exponents and type of walls having c fixed at 10. Observing Fig. 4, where $\gamma = 3$, one may notice the "so-called" knees in the form of the densities. These bending points are not artifacts of the simulation study but are inherent form of the density corresponding to transition from small path loss exponent ($\gamma = 2$ and slightly greater) resulting in Gamma-like unimodal densities to high exponents around 4 characterized by distributions with no mode. Observe that the bending point gets smoother when the material of walls becomes harder to penetrate for electromagnetic radiation. For $\gamma = 4$ the densities are completely smooth implying that the wall material effect is similar to the increase of the path loss exponent.

The D2D case for different propagation loss exponents and $c = 10$ is shown in Fig. 6. For D2D scenario the form of the densities corresponds to $\gamma = 3$. As one may observe increasing the propagation losses caused by walls materials the mass of the density tends to concentrate on the left hand side making the tail of the distribution lighter. Thus, the propagation conditions become better.

Finally, the performance of the D2D connectivity for a number of different wall materials and path loss exponents is illustrated in Fig. 7. Here, the results are presented in dB using the transformation $10 \log_{10} I$. The data for propagation loss for different materials is taken from [14].

CONCLUSIONS

In this paper we analyzed performance of densely deployed indoor wireless systems for both D2D and AP configurations. For uplink AP configuration we obtained analytical results and demonstrated that for downlink AP and D2D configurations the interference path are, in fact, dependent preventing mathematical analysis in these cases. These scenarios have been addressed using simulation study.

Our results indicate that the performance of the considered scenarios is affected by the interplay between the propagation loss exponent and the types of walls between rooms. We evaluated the defined scenarios for a range of these parameters.

ACKNOWLEDGMENTS

The reported study was partially supported by RFBR, research projects No. 14-07-00090, 15-07-03051.

REFERENCES

- [1] Ericsson, AB, "Ericsson mobility report," 2015.
- [2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," 2015.
- [3] J. Andrews, H. Claussen, M. Dohler, and S. Rangan, "Femtocells: Past, present, and future," *IEEE JSAC*, vol. 30, pp. 497–508, Apr. 2012.
- [4] J. Lee, H. Wang, J. Andrews, and D. Hong, "Outage probability of cognitive relay networks with interference constraints," *IEEE Trans. Wir. Comm.*, vol. 10, pp. 390–395, Feb. 2011.
- [5] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmı, "Device-to-device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510–1520, Apr. 2014.

- [6] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.
- [7] V. Sathya, A. Ramamurthy, S. Kumar, and B. Tamma, "On improving SINR in LTE hetnets with D2D relays," *Computer Communications*, 2015.
- [8] J. G. Andrews, R. K. Ganti, M. Haenggi, N. Jindal, and S. Weber, "A primer on spatial modeling and analysis in wireless networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 156–163, 2010.
- [9] D. Moltchanov, "Survey paper: Distance distributions in random networks," *Ad Hoc Netw.*, vol. 10, no. 6, pp. 1146–1166, 2012.
- [10] L. Zhou, "D2D communication meets big data: From theory to application," *Mobile Networks and Applications*, vol. 20, no. 6, pp. 783–792, 2015.
- [11] A. Samuylov, A. Ometov, V. Begishev, R. Kovalchukov, D. Moltchanov, Y. Gaidamaka, K. Samouylov, S. Andreev, and Y. Koucheryavy, "Analytical performance estimation of network-assisted D2D communications in urban scenarios with rectangular cells," *Transactions on Emerging Telecommunications Technologies*, 2015.
- [12] B. Levin, *Theoretical Foundations of Statistical Radio Engineering*. M.: Radio i Svyaz, 3rd ed., 1989.
- [13] ITU-R, "Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz," *Recommendation ITU-R P.1238-8*, 07-2015.
- [14] www.digi.com, "Indoor path loss," application note, accessed on: 12.01.2016, Digi, 2012.

STOCHASTIZATION OF ONE-STEP PROCESSES IN THE OCCUPATIONS NUMBER REPRESENTATION

Anna V. Korolkova, Ekaterina. G. Eferina,
Eugeniy B. Laneev

Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia,
Email: akorolkova@sci.pfu.edu.ru, eg.eferina@gmail.com,
laneev_eb@pfur.ru

Irina A. Gudkova

Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
and Institute of Informatics Problems, FRC CSC RAS,
44-2 Vavilova Str., Moscow, 119333, Russia,
Email: igudkova@sci.pfu.edu.ru

Leonid A. Sevastianov

Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
and Bogoliubov Laboratory of Theoretical Physics
Joint Institute for Nuclear Research
Joliot-Curie 6, Dubna, Moscow region, 141980, Russia,
Email: leonid.sevast@gmail.com

Dmitry S. Kulyabov

Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
and Laboratory of Information Technologies
Joint Institute for Nuclear Research
Joliot-Curie 6, Dubna, Moscow region, 141980, Russia,
Email: yamadharna@gmail.com

KEYWORDS

Occupation numbers representation, Fock space, Dirac notation, one-step processes, stochastic differential equations, master equation

ABSTRACT

By the means of the method of stochastization of one-step processes we get the simplified mathematical model of the original stochastic system. We can explore these models by standard methods, as opposed to the original system. The process of stochastization depends on the type of the system under study. We want to get a unified abstract formalism for stochastization of one-step processes. This formalism should be equivalent to the previously introduced. To implement an abstract approach we use the representation of occupation numbers. In this presentation we use the operator formalism. A feature of this formalism is the use of abstract linear operators which are independent from the state vectors. We use the formalism of Green's functions in order to deal with operators. We get a fully coherent formalism by using the occupation numbers representation. With its help we can get simplified stochastic model of the original system. We demonstrate the equivalence of the occupation number representation and the state vectors representation by using a one-step process example. We have suggested a convenient formalism for unified description of stochastic systems. Also, this method can be extended for the study of nonlinear stochastic systems.

INTRODUCTION

When modeling various physical and technical systems, we often can model them in the form of a one-step processes (see Demidova et al. (2013, 2014); Velieva et al. (2014); Basharin et al. (2009)). Then there is the problem of adequate representation and study of the resulting model. The formalism of stochastization of one-step processes has been developed by our group for quite a long time. But so far, our efforts have been aimed at getting more models than on their investigation. For the statistical systems in addition to representation of the state vectors (combinatorial approach) the representation of the occupation numbers (operator approach) (see Hnatič et al. (2016); Grassberger and Scheunert (1980); Täuber (2005); Janssen and Täuber (2005); Mabilia et al. (2006)) is also used. This representation is especially well suited for the system with a variable number of elements description. In addition, for this representation there are effective methods for solving equations based on the formalism of Green's functions and perturbation theory.

In this paper, we want to demonstrate the methodology of both approaches.

The structure of the article is as follows. In the first section basic notations and conventions are introduced. The ideology of the method of stochastization of one-step process and its components are described in the second section. Then the interaction schemes and master equation overview are presented in the next section. The combinatorial method of modelling is discussed in the following section. The operator model approach is presented in the last section, where, in

particular, the algorithm of transition to the occupation number representation is described.

NOTATIONS AND CONVENTIONS

- 1) The abstract indices notation (see Penrose and Rindler (1987)) is used in this work. Under this notation a tensor as a whole object is denoted just as an index (e.g., x^i), components are denoted by underlined index (e.g., $x^{\underline{i}}$).
- 2) We will adhere to the following agreements. Latin indices from the middle of the alphabet (i, j, k) will be applied to the space of the system state vectors. Latin indices from the beginning of the alphabet (a) will be related to the Wiener process space. Greek indices (α) will set a number of different interactions in kinetic equations.

GENERAL REVIEW OF THE METHODOLOGY

Our methodology is completely formalized in such a way that it is sufficient when the original problem is formulated accordingly. It should be noted that the most of the models under our study can be formalized as a one-step process (see van Kampen (2011); Gardiner (1985)). In fact, for this type of models we developed this methodology, but it may be expanded for other processes.

First we transform our model to the one-step process (see Fig. 1). Next, we need to formalize this process in the form of interaction schemes ¹ (see Demidova et al. (2013, 2014); Hnatič et al. (2016)).

Each scheme has its own interaction semantics. Semantics leads directly to the master equation (see van Kampen (2011); Gardiner (1985)). However, the master equation has usually rather complex structure that makes it difficult for direct study and solution. Our technique involves two possibilities (see Fig. 2):

- computational approach — the solution of the master equation with help of perturbation theory;
- modeling approach — the approximate models are obtained in the form of Fokker–Planck and Langevin equations.

The computational approach allows to obtain a concrete solution for the studied model. In our methodology, this approach is associated with perturbation theory (see Hnatič et al. (2013); Hnatič and Honkonen (2000); Hnatič et al. (2011)). Methodologically, this method is quite simple. Each expansion element appears in the form of Feynman diagrams. However, with increasing order of the expansion, the number of Feynman diagrams increases rapidly and can reach tens or hundreds of thousands. It is quite natural that this should involve high-performance computing.

The model approach provides a model that is convenient to study numerically and qualitatively. In addition, this approach assumes the iterative process of research: the obtained approximate model can be specified and changed, which leads to the correction of initial interaction schemes.

¹The analogs of the interaction schemes are the equations of chemical kinetics, reaction particles and etc.

In this article we will describe the model approach.

There are two ways of building the master equation²

- combinatorial approach (see Fig. 3);
- operator approach (see Fig. 4).

In the combinatorial approach, all operations are performed in the space of states of the system, so we deal with a particular system throughout manipulations with the model.

For the operator approach we can abstract from the specific implementation of the system under study. We are working with abstract operators. We return to the state space only at the end of the calculations. In addition, we choose a particular operator algebra on the basis of symmetry of the problem.

These two approaches are belong to different paradigms of physical theories construction. Accordingly, they are complementary. Some constructions are simpler in one approach, others are simpler in another. For example, in the combinatorial approach, the process of obtaining the approximate models is more convenient, but not more easy. For perturbation theory it is easier to expand in a series the master equation by using the operator formalism. In addition, the operator formalism is suitable for describing the transient processes and non-stationary statistical systems.

Interaction schemes

The system state is defined by the vector $\varphi^i \in \mathfrak{R}^n$, where n is system dimension ³. The operator $I_j^i \in \mathfrak{N}_0^n \times \mathfrak{N}_0^n$ describes the state of the system before the interaction, the operator $F_j^i \in \mathfrak{N}_0^n \times \mathfrak{N}_0^n$ describes the state of the system after the interaction⁴. The result of interaction is the system transition from one state to another one.

There are s types of interaction in our system, so instead of I_j^i and F_j^i operators we will use operators $I_j^{i\alpha} \in \mathfrak{N}_0^n \times \mathfrak{N}_0^n \times \mathfrak{N}_+^s$ and $F_j^{i\alpha} \in \mathfrak{N}_0^n \times \mathfrak{N}_0^n \times \mathfrak{N}_+^s$.

The interaction of the system elements will be described by interaction schemes, which are similar to schemes of chemical kinetics Waage and Gulberg (1986); Gorban and Yablonsky (2015):

$$I_j^{i\alpha} \varphi^j \xrightleftharpoons[\underline{-k_\alpha}]{+k_\alpha} F_j^{i\alpha} \varphi^j, \quad \underline{\alpha} = \overline{1, s}, \quad (1)$$

the Greek indices specify the number of interactions and Latin are the system order. The coefficients $+k_\alpha$ and $-k_\alpha$ have meaning intensity (speed) of interaction.

The state transition is given by the operator:

$$r_j^{i\alpha} = F_j^{i\alpha} - I_j^{i\alpha}. \quad (2)$$

²In quantum field theory the path integrals approach can be considered as an analogue of the combinatorial approach and the method of second quantization as analog of the operator approach.

³For brevity, we denote the module over the field \mathbb{R} just as \mathfrak{R} . Accordingly, \mathfrak{N} , \mathfrak{N}_0 , \mathfrak{N}_+ are modules over rings \mathbb{N} , \mathbb{N}_0 (cardinal numbers with 0), \mathbb{N}_+ (cardinal numbers without 0).

⁴The component dimension indices take on values $\underline{i}, \underline{j} = \overline{1, n}$

⁵The component indices of number of interactions take on values $\underline{\alpha} = \overline{1, s}$

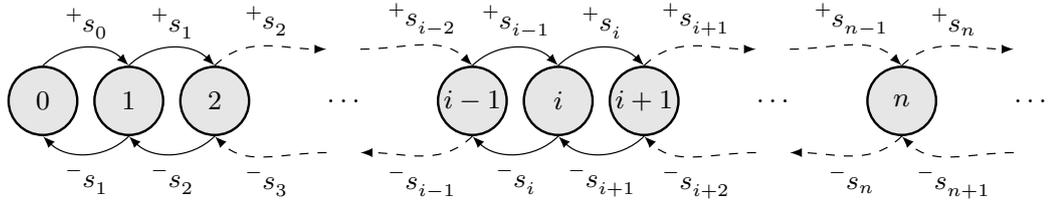


Fig. 1. One-step process

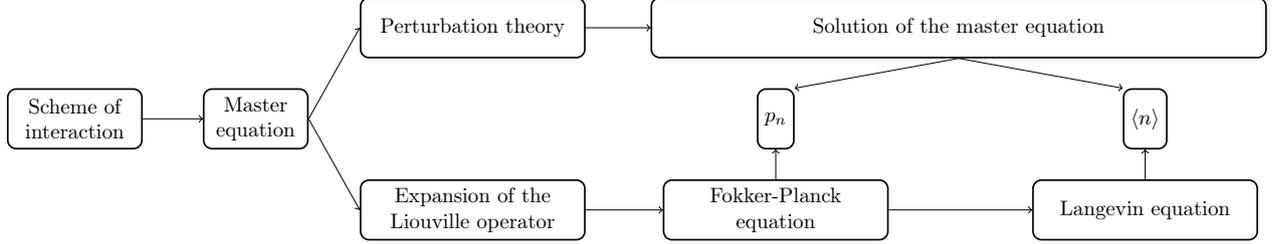


Fig. 2. The general structure of the methodology

Thus, one step interaction $\underline{\alpha}$ in forward and reverse directions can be written as

$$\begin{aligned}\varphi^i &\rightarrow \varphi^i + r_j^{i\alpha} \varphi^j, \\ \varphi^i &\rightarrow \varphi^i - r_j^{i\alpha} \varphi^j.\end{aligned}$$

We can also write (1) not in the form of vector equations but in the form of sums:

$$I_j^{i\alpha} \varphi^j \delta_i \xrightleftharpoons[-k_\alpha]{+k_\alpha} F_j^{i\alpha} \varphi^j \delta_i,$$

where $\delta_i = (1, \dots, 1)$.

Also the following notation will be used:

$$I^{i\alpha} := I_j^{i\alpha} \delta^j, \quad F^{i\alpha} := F_j^{i\alpha} \delta^j, \quad r^{i\alpha} := r_j^{i\alpha} \delta^j.$$

The master equation

For the system description we will use the master equation,⁶ which describes the transition probabilities for Markov process (see van Kampen (2011); Gardiner (1985)):

$$\frac{\partial p(\varphi_2, t_2 | \varphi_1, t_1)}{\partial t} = \int [w(\varphi_2 | \psi, t_2) p(\psi, t_2 | \varphi_1, t_1) - w(\psi | \varphi_2, t_2) p(\varphi_2, t_2 | \varphi_1, t_1)] d\psi,$$

where $w(\varphi | \psi, t)$ is the probability of transition from the state ψ to the state φ for unit time.

Fixing the initial values of φ_1, t_1 , we can write the equation for subensemble:

$$\frac{\partial p(\varphi, t)}{\partial t} = \int [w(\varphi | \psi, t) p(\psi, t) - w(\psi | \varphi, t) p(\varphi, t)] d\psi. \quad (3)$$

⁶Master equation can be considered as an implementation of the Kolmogorov equation. However, the master equation is more convenient and has an immediate physical interpretation (see van Kampen (2011)).

If a domain of φ is a discrete one then the (3) can be written as follows (the states are numbered by n and m):

$$\frac{\partial p_n(t)}{\partial t} = \sum_m [w_{nm} p_m(t) - w_{mn} p_n(t)], \quad (4)$$

where the p_n is the probability of the system to be in a state n at time t , w_{nm} is the probability of transition from the state m into the state n per unit time.

There are two types of system transition from one state to another (based on one-step processes) as a result of system elements interaction: in the forward direction ($\varphi^i + r_j^{i\alpha} \varphi^j$) with the probability $+s_\alpha(\varphi^k)$ and in the opposite direction ($\varphi^i - r_j^{i\alpha} \varphi^j$) with the probability $-s_\alpha(\varphi^k)$ (fig. 1). The matrix of transition probabilities has the form:

$$w_\alpha(\varphi^i | \psi^i, t) = +s_\alpha \delta_{\varphi^i, \psi^{i+1}} + -s_\alpha \delta_{\varphi^i, \psi^{i-1}}, \quad \underline{\alpha} = \overline{1, s},$$

where $\delta_{i,j}$ is Kronecker delta.

Thus, the general form of the master equation for the state vector φ^i , changing by steps with length $r_j^{i\alpha} \varphi^j$, is:

$$\begin{aligned}\frac{\partial p(\varphi^i, t)}{\partial t} &= \sum_{\underline{\alpha}=1}^s \left\{ -s_\alpha(\varphi^i + r^{i\alpha}, t) p(\varphi^i + r^{i\alpha}, t) + \right. \\ &\quad \left. + s_\alpha(\varphi^i - r^{i\alpha}, t) p(\varphi^i - r^{i\alpha}, t) - \right. \\ &\quad \left. - [s_\alpha(\varphi^i) + -s_\alpha(\varphi^i)] p(\varphi^i, t) \right\}. \quad (5)\end{aligned}$$

COMBINATORIAL APPROACH

We will obtain the function $+s_\alpha$ and $-s_\alpha$ for equation (5) with use of combinatorial approach.

The transition probabilities

The transition rates $+s_\alpha$ and $-s_\alpha$ are proportional to the number of ways of choosing the number of arrangements of

φ^i to $I^{i\alpha}$ (denoted as $A_{\varphi^i}^{i\alpha}$) and to $F^{i\alpha}$ (denoted as $A_{\varphi^i}^{F^{i\alpha}}$) and defined by:

$$\begin{aligned} +s_{\underline{\alpha}} &= +k_{\underline{\alpha}} \prod_{i=1}^n A_{\varphi^i}^{i\alpha} = +k_{\underline{\alpha}} \prod_{i=1}^n \frac{\varphi^i!}{(\varphi^i - I^{i\alpha})!}, \\ -s_{\underline{\alpha}} &= -k_{\underline{\alpha}} \prod_{i=1}^n A_{\varphi^i}^{F^{i\alpha}} = -k_{\underline{\alpha}} \prod_{i=1}^n \frac{\varphi^i!}{(\varphi^i - F^{i\alpha})!}. \end{aligned} \quad (6)$$

Replacing in (6) the $\varphi(\varphi - 1) \cdots (\varphi - (n - 1))$ -type combinations on $(\varphi)^n$ we obtain for Fokker–Planck equation⁷:

$$\begin{aligned} +s_{\underline{\alpha}} &= +k_{\underline{\alpha}} \prod_{i=1}^n (\varphi^i)^{i\alpha}, \\ -s_{\underline{\alpha}} &= -k_{\underline{\alpha}} \prod_{i=1}^n (\varphi^i)^{F^{i\alpha}}. \end{aligned}$$

Fokker–Planck equation

The Fokker–Planck equation is a special case of the master equation and can be regarded as its approximation. We can get through the expansion of the master equation in a series up to the second order. We will use the decomposition of the Kramers–Moyal (see Gardiner (1985)) (for simplicity it is written for the one-dimensional case):

$$\frac{\partial p(\varphi, t)}{\partial t} = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial \varphi^n} [\xi^n(\varphi) p(\varphi, t)],$$

where

$$\xi^n(\varphi) = \int_{-\infty}^{\infty} (\psi - \varphi)^n w(\psi|\varphi) d\psi.$$

By dropping the terms with order higher than the second one, we obtain the Fokker–Planck equation:

$$\frac{\partial p(\varphi, t)}{\partial t} = -\frac{\partial}{\partial \varphi} [A(\varphi)p(\varphi, t)] + \frac{\partial^2}{\partial \varphi^2} [B(\varphi)p(\varphi, t)],$$

and for multivariate case

$$\begin{aligned} \frac{\partial p(\varphi^k, t)}{\partial t} &= -\frac{\partial}{\partial \varphi^i} [A^i(\varphi^k)p(\varphi^k, t)] + \\ &+ \frac{1}{2} \frac{\partial^2}{\partial \varphi^i \partial \varphi^j} [B^{ij}(\varphi^k)p(\varphi^k, t)], \end{aligned} \quad (7)$$

where

$$\begin{aligned} A^i &:= A^i(\varphi^k) = r^{i\alpha} \left[+s_{\underline{\alpha}} - -s_{\underline{\alpha}} \right], \\ B^{ij} &:= B^{ij}(\varphi^k) = r^{i\alpha} r^{j\alpha} \left[+s_{\underline{\alpha}} - -s_{\underline{\alpha}} \right]. \end{aligned} \quad (8)$$

As can be seen from the (8), the coefficients of the Fokker–Planck equation can be obtained directly from the (2) and (6), that is, in this case, it is not necessary to write down the master equation.

⁷This change corresponds to a series expansion.

Langevin equation

The Langevin equation which corresponds to the Fokker–Planck equation:

$$d\varphi^i = a^i dt + b_a^i dW^a, \quad (9)$$

where $a^i := a^i(\varphi^k)$, $b_a^i := b_a^i(\varphi^k)$, $\varphi^i \in \mathfrak{R}^n$ is the system state vector, $W^a \in \mathbb{R}^m$ is the m -dimensional Wiener process⁸. Latin indices from the middle of the alphabet will be applied to the system state vectors (the dimensionality of space is n), and Latin indices from the beginning of the alphabet denote the variables related to the Wiener process vector (the dimensionality of space is $m \leq n$).

The connection between the equations (7) and (9) is expressed by the following relationships:

$$A^i = a^i, \quad B^{ij} = b_a^i b^{ja}.$$

It can be seen that the second term of the Langevin equation is a square root, which has a complicated form in the multidimensional case. However, we note that is often used is the square of the second term of the Langevin equation, so often calculate the root is not required.

OPERATOR APPROACH

Occupation numbers representation

Occupation number representation is the main language in the description of many-body physics. The main elements of the language are the wave functions of the system with information about how many particles are in each single-particle state. The creation and annihilation operators are used for system states change. The advantages of this formalism are following:

- it is possible to consider systems with a variable number of particles (non-stationary systems);
- system statistics (Fermi–Dirac or Bose–Einstein) is automatically included in the commutation rules for the creation and annihilation operators;
- this is the second major formalism (along with the path integral) for the quantum perturbation theory description.

The method of application of the formalism of second quantization for the non-quantum systems (statistical, deterministic systems) was studied in a series of articles (see Doi (1976a,b); Grassberger and Scheunert (1980); Peliti (1985)).

The Dirac notation is commonly used for occupation numbers representation recording.

Dirac notation

This notation is proposed by P. A. M. Dirac⁹ (see Dirac (1939)). Under this notation, state of the system is described by an element of the projective Hilbert space \mathcal{H} . The vector

⁸the Wiener process is realized as $dW = \varepsilon \sqrt{dt}$, where $\varepsilon \sim N(0, 1)$ — the normal distribution with mean 0 and variation 1.

⁹The notation is based on the notation, proposed by G. Grassmann in 1862 (see (Cajori, 1929, p. 134)).

$\varphi^i \in \mathcal{H}$ is defined as $|i\rangle$, and covariant vector (covector) $\varphi_i \in \mathcal{H}^* := \mathcal{H}_\bullet$ is defined as $\langle i|$. Conjunction operation is used for raising and lowering of indices¹⁰:

$$\varphi_i^* := \varphi_i = (\varphi^i)^\dagger \equiv \langle i| = |i\rangle^\dagger. \quad (10)$$

The scalar product is as follows:

$$\varphi_i \varphi^i \equiv \langle i|i\rangle.$$

The tensor product is:

$$\varphi_j \varphi^i \equiv |i\rangle \langle j|.$$

However, this notation (10) is normally used for some dedicated vectors, such as basis (δ_i^j) or eigenvectors. Then conventional vectors using the notation of the following form:

$$|\varphi\rangle \equiv \varphi^i, \quad \langle i|\varphi\rangle \equiv \varphi^i \delta_i^i = \varphi^i.$$

Linear operators are usually used for operations with vectors. Let $A_j^i : \mathcal{H}^\bullet \rightarrow \mathcal{H}_\bullet$ be a linear operator. Then the inner product will be as follows:

$$A_j^i \varphi_i \psi^j \equiv \langle \varphi|A|\psi\rangle.$$

Component representation of a linear operator can be written as follows:

$$A_j^i = A_j^i \delta_i^i \delta_j^j \equiv \langle i|A|j\rangle.$$

Creation and annihilation operators

The transition to the space of occupation numbers is not a unitary transformation. However, the algorithm of transition (specific to each task) can be constructed.

Let's write the master equation (4) in the occupation number representation. We will consider a system that does not depend on the spatial variables. For simplicity, we consider the one-dimensional version.

Let's denote in (4) the probability that there are n particles in our system as φ_n :

$$\varphi_n := p_n(\varphi, t).$$

The vector space \mathcal{H} consists of states of φ .

We introduce a scalar product, exclusive ($\langle \rangle_{\text{ex}}$) and inclusive ($\langle \rangle_{\text{in}}$). Let $|n\rangle$ are basis vectors.

$$\langle \varphi|\psi\rangle_{\text{ex}} = \sum_n n! p_n^*(\varphi) p^n(\psi); \quad (11)$$

$$\langle \varphi|\psi\rangle_{\text{in}} = \sum_n \frac{1}{k!} n_k^*(\varphi) n^k(\psi). \quad (12)$$

¹⁰In this case, we use Hermitian conjugation \bullet^\dagger . The sign of the complex conjugate \bullet^* in this entry is superfluous.

There n_k are factorial moments:

$$n_k(\varphi) = \langle n(n-1) \cdots (n-k+1) \rangle = \frac{\partial^k}{\partial z^k} G(z, \varphi)|_{z=1}, \quad (13)$$

G is generating function:

$$G(z, \varphi) = \sum_n z^n p_n(\varphi). \quad (14)$$

Normalizing for the generating function is obvious. Let $z = 1$ in equation (14). Then

$$\sum_n p_n(\varphi) = 1, \quad G(1, \varphi) = 1, \quad n_0(\varphi) = 1.$$

From $p_n(m) = \delta_n^m$ and (11) we can obtain:

$$\langle n|m\rangle_{\text{ex}} = n! \delta_n^m. \quad (15)$$

The state vectors:

$$|\varphi\rangle = \sum_n p_n(\varphi) |n\rangle = \sum_n \varphi_n |n\rangle =: \varphi_n |n\rangle. \quad (16)$$

In view of (15) the following expression may be written:

$$\varphi_n = \frac{1}{n!} \langle n|\varphi\rangle_{\text{ex}}. \quad (17)$$

Let's use creation and annihilation operators:

$$\begin{aligned} \pi |n\rangle &= |n+1\rangle, \\ a |n\rangle &= n |n-1\rangle \end{aligned}$$

and commutation rule¹¹:

$$[a, \pi] = 1. \quad (18)$$

If the form of scalar product is (11) then with the help of (18) it is obviously that the our system is described by Bose–Einstein statistics.

From the relation (15) we obtain:

$$\langle m|a^\dagger|n\rangle_{\text{ex}} = \langle m|\pi|n\rangle_{\text{ex}},$$

and for the scalar product (11) the following statement is valid:

$$a^\dagger = \pi.$$

Now we can write the inclusive scalar product from(14) and (13):

$$\begin{aligned} p_n(\varphi) &= \frac{1}{n!} \langle n|\varphi\rangle_{\text{ex}} = \frac{1}{n!} \langle 0|a^n|\varphi\rangle_{\text{ex}}, \\ n_k(\varphi) &= \langle 0|\exp(a)a^k|\varphi\rangle_{\text{ex}}. \end{aligned} \quad (19)$$

Taking into account the (12), we obtain the inclusive scalar product

$$\langle \varphi|\psi\rangle_{\text{in}} = \langle \varphi|\exp(\pi) \exp(a)|\psi\rangle_{\text{ex}}.$$

Then the expressions (19) will take the following form:

$$\begin{aligned} p_n(\varphi) &= \frac{1}{n!} \langle 0|\exp(-a)a^n|\varphi\rangle_{\text{in}}, \\ n_k(\varphi) &= \langle 0|a^k|\varphi\rangle_{\text{in}} = \langle k|\varphi\rangle_{\text{in}}. \end{aligned}$$

¹¹In fact, $a\pi|n\rangle - \pi a|n\rangle = (n+1)|n\rangle - n|n\rangle = |n\rangle$.

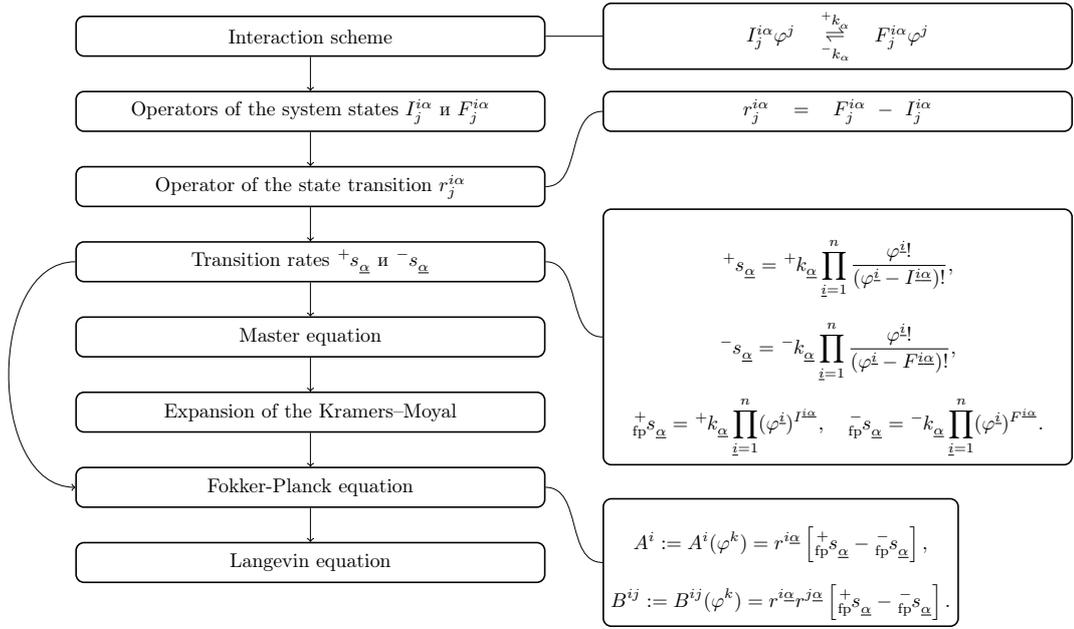


Fig. 3. Combinatorial modeling approach

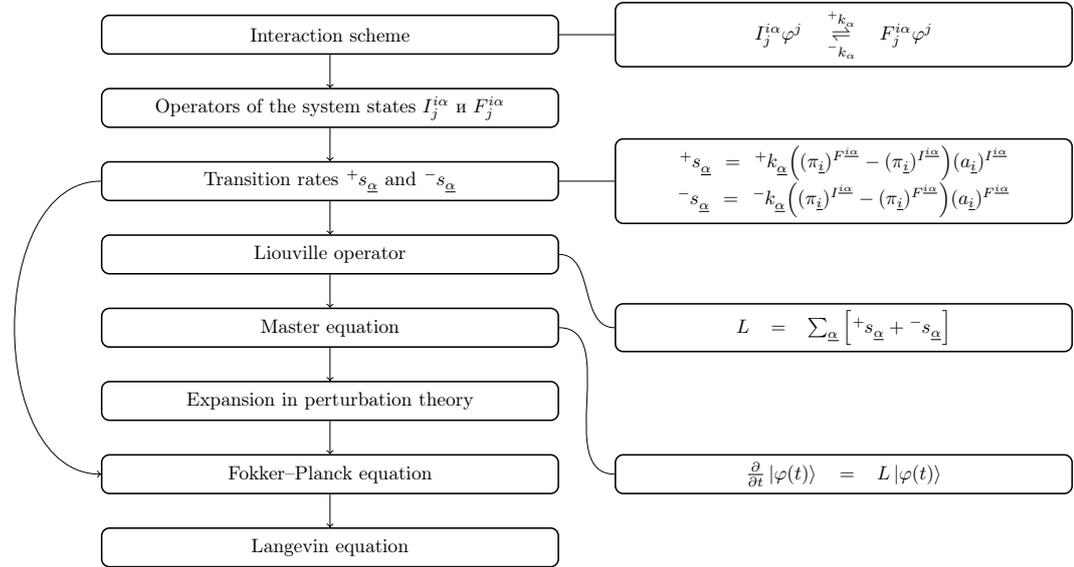


Fig. 4. Operator modeling approach

The Liouville operator

In the occupation numbers formalism the master equation becomes the Liouville equation:

$$\frac{\partial}{\partial t} |\varphi(t)\rangle = L |\varphi(t)\rangle. \quad (20)$$

The Liouville operator L satisfies the relation:

$$\langle 0 | L = 0.$$

From (4), (16), (17) and (20) we obtain:

$$\frac{\partial p_n}{\partial t} = \frac{1}{n!} \left\langle n \left| \frac{\partial}{\partial t} \right| \varphi \right\rangle = \frac{1}{n!} \langle n | L | \varphi \rangle \equiv$$

$$\equiv \sum_m [w_{nm} p_m - w_{mn} p_n],$$

The Liouville equation (20) in the form of a single equation writes down the master equations (4) for different values of n .

The following Liouville operator corresponds to the scheme (1):

$$L = \sum_{\alpha, \underline{i}} \left[+k_{\alpha} \left((\pi_{\underline{i}})^{F^{i\alpha}} - (\pi_{\underline{i}})^{I^{i\alpha}} \right) (a_{\underline{i}})^{I^{i\alpha}} + \right. \\ \left. + -k_{\alpha} \left((\pi_{\underline{i}})^{I^{i\alpha}} - (\pi_{\underline{i}})^{F^{i\alpha}} \right) (a_{\underline{i}})^{F^{i\alpha}} \right]. \quad (21)$$

As we can see, in the case of the occupation numbers

formalism (21) the schemes of interaction encode the system under study in more universal and transparent way than in the representation of the state vectors (6).

CONCLUSIONS

We presented a preliminary overview of the use of the occupation numbers representation for the record of models describing stochastic statistical systems. However, the model approach does not always present the advantages of described method. We assume that the advantages of the method will reveal most explicitly when computation approach is used.

a) Notes and Comments: The work is partially supported by RFBR grants No's 14-01-00628, 15-07-08795, and 16-07-00556.

REFERENCES

- Basharin, G. P., Samouylov, K. E., Yarkina, N. V. and Gudkova, I. A. (2009), A new stage in mathematical teletraffic theory, *Automation and Remote Control* 70(12), 1954–1964.
- Cajori, F. (1929), *A History of Mathematical Notations*, Vol. 2.
- Demidova, A. V., Korolkova, A. V., Kulyabov, D. S. and Sevastianov, L. A. (2013), The method of stochastization of one-step processes, *Mathematical Modeling and Computational Physics*, JINR, Dubna, p. 67.
- Demidova, A. V., Korolkova, A. V., Kulyabov, D. S. and Sevastianov, L. A. (2014), The method of constructing models of peer to peer protocols, 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, pp. 557–562.
- Dirac, P. A. M. (1939), A new notation for quantum mechanics, *Mathematical Proceedings of the Cambridge Philosophical Society* 35(03), 416.
- Doi, M. (1976a), Second quantization representation for classical many-particle system, *Journal of Physics A: Mathematical and General* 9(9), 1465–1477.
- Doi, M. (1976b), Stochastic theory of diffusion-controlled reaction, *Journal of Physics A: Mathematical and General* 9(9), 1479–1495.
- Gardiner, C. W. (1985), *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences*, Springer Series in Synergetics.
- Corban, A. N. and Yablonsky, G. S. (2015), Three Waves of Chemical Dynamics, *Math. Model. Nat. Phenom. Vol.* 10(5), 1–5.
- Grassberger, P. and Scheunert, M. (1980), Fock-Space Methods for Identical Classical Objects, *Fortschritte der Physik* 28(10), 547–578.
- Hnatič, M., Eferina, E. G., Korolkova, A. V., Kulyabov, D. S. and Sevastianov, L. A. (2016), Operator Approach to the Master Equation for the One-Step Process, *EPJ Web of Conferences* 108, 02027.
- Hnatič, M., Honkonen, J. and Lučivjanský, T. (2013), Field-theoretic technique for irreversible reaction processes, *Physics of Particles and Nuclei* 44(2), 316–348.
- Hnatič, M. and Honkonen, J. (2000), Velocity-fluctuation-induced anomalous kinetics of the $A+A \rightarrow \zeta$ reaction, *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* 61(4 Pt A), 3904–3911.
- Hnatič, M., Honkonen, J. and Lučivjanský, T. (2011), Field theory approach in kinetic reaction: Role of random sources and sinks, *Theoretical and Mathematical Physics* 169(1), 1489–1498.
- Janssen, H.-K. and Täuber, U. C. (2005), The field theory approach to percolation processes, *Annals of Physics* 315(1), 147–192.
- Mobilia, M., Georgiev, I. T. and Täuber, U. C. (2006), Fluctuations and correlations in lattice models for predator-prey interaction, *Physical Review E* 73(4), 040903.

- Peliti, L. (1985), Path integral approach to birth-death processes on a lattice, *Journal de Physique* 46(9), 1469–1483.
- Penrose, R. and Rindler, W. (1987), *Spinors and Space-Time: Volume I, Two-Spinor Calculus and Relativistic Fields*, Vol. 1, Cambridge University Press.
- Täuber, U. C. (2005), *Field-Theory Approaches to Nonequilibrium Dynamics, Ageing and the Glass Transition*, Vol. 716, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 295–348.
- van Kampen, N. G. (2011), *Stochastic Processes in Physics and Chemistry*, North-Holland Personal Library, Elsevier Science.
- Velieva, T. R., Korolkova, A. V. and Kulyabov, D. S. (2014), Designing installations for verification of the model of active queue management discipline RED in the GNS3, 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, pp. 570–577.
- Waage, P. and Gulberg, C. M. (1986), Studies concerning affinity, *J. Chem. Educ.* 63(12), 1044.

AUTHOR BIOGRAPHIES

ANNA V. KOROLKOVA received his Ph.D. in Mathematics in 2010. Since then, she has worked as associate professor in Peoples' Friendship University of Russia. Her current research activity focuses on mathematical modeling. Her email address is akorolkova@sci.pfu.edu.ru.

DMITRY S. KULYABOV received his Ph.D. in Physics in 2000. Since then, he has worked as associate professor in Peoples' Friendship University of Russia. His current research activity focuses on mathematical modeling. His email address is yamadharma@gmail.com.

LEONID A. SEVASTIANOV received his D.Sc. in Phys.-Math. in 1999. Since then, he has worked as full professor in Peoples' Friendship University of Russia. His current research activity focuses on mathematical modeling. His email address is leonid.sevast@gmail.com.

EKATERINA G. EFERINA graduate student in Peoples' Friendship University of Russia. Her current research activity focuses on mathematical modeling. Her email address is eg.eferina@gmail.com.

IRINA A. GUDKOVA received his Ph.D. in Mathematics in 2013. Since then, he has worked as associate professor in Peoples' Friendship University of Russia. Her current research activity focuses on queuing theory. Her email address is igudkova@sci.pfu.edu.ru.

EUGENY B. LANEV received his D.Sc. in Mathematics in 2005. Since then, he has worked as full professor in Peoples' Friendship University of Russia. His current research activity focuses on mathematical modeling. His email address is lanev_eb@pfur.ru.

ON ANALYTICAL MODELING OF IMS CONFERENCING SERVER

Pavel Abaev
Vitaly Beschastny
Alexey Tsarev

Department of Applied Probability and Informatics
Peoples' Friendship University of Russia
Mikluho-Maklaya str., 6
Moscow 117198, Russia
E-mail: {pabaev, vbeschastny, atsarev}@sci.pfu.edu.ru

KEYWORDS

IMS, SIP, conferencing server, exhaustive service, vacations.

ABSTRACT

The IP Multimedia Subsystem is an architectural framework for delivering multimedia services over an Internet Protocol (IP) network. Originally, it was specified for wireless networks, but has since evolved to incorporate fixed line access as well. It forms part of a Next Generation Network (NGN) which is defined as a packet-based network where the service functionality is independent of the underlying transport technologies. This allows new converged services to be implemented on top of an existing packet switched network.

The centralized conferencing framework limits the ability to have large number of participants because of the overhead on the server. One of the possible efficient solution for increasing of server's performance is grouping meeting requests the same video conference and serving them at one time on an edge-proxy. The queuing model with batch exhaustive service and working vacations is constructed and analyzed. The stationary distribution and the formulas for main performance measurements are obtained. In addition, the optimization problem for increasing a server performance is formulated.

INTRODUCTION

The IP Multimedia Subsystem is an architectural framework for delivering multimedia services over an Internet Protocol (IP) network. IMS uses SIP protocol to support communication sessions with multiple participants (Camarillo, 2008; TS 23.228, 2004). SIP is an application-layer signaling protocol for creating, modifying, and terminating sessions with one or more participants. In 1996, Henning Schulzrinne and Mark Handley started working on creating a SIP protocol within the IETF (Internet Engineering Task Force) project for developing a series of protocols for the

provision of multimedia services. In November 2000, SIP was accepted as a 3GPP signaling protocol and main protocol of the IMS architecture (TS.24.229, 2015). In 2002, the RFC3261 (Rosenberg, 2002) recommendation which determines the current protocol form was accepted.

SIP supports both centralized and decentralized frameworks of multi-party communication. Overview of the most recent researches done by Mishra in (Mishra, 2014) shows that distributed approach allows to enhance the conferencing server capacity but it has some limitations with full security and control over the complete network due to unavailability of central server; it increases signaling delay due to multiple functional entities participating in conference session setup. In centralized conferencing architecture (Tien, 2010), a single User Agent (UA) or focus has a direct relation with each participant. Therefore, it shares many problems when the number of participants in conference increases. The centralized conferencing framework does not meet the increasing requirements for server's capacity due to growing demand for conferencing service. However, both of these approaches do not take into account SIP server overloading problems (Abaev, 2012; Abaev 2013). The SIP main shortcomings concerning overload prevention described in (RFC 5390, 2008) by Rosenberg.

One of the efficient approach to enhance server capability is to improve the way how the server serves incoming conferencing request. We consider the method of grouping requests related to the same conference and serving them at one time. This approach will reduce session establishment time and will increase the server throughput.

This paper is organized as follows. We analyze 3GPP standards for basic call flow for conference service session establishment. Then we investigate message grouping approach on the edge server and construct a mathematical model in the form of queuing system with the buffer of finite capacity. We obtain formulas for the main performance measures and perform numerical experiments.

IMS CONFERENCING SERVICE CALL FLOW WITH REQUEST/RESPONSE GROUPING APPROACH

SIP is a request/response-based protocol. According to (TS.24.229, 2015), IMS architecture consists of the following components:

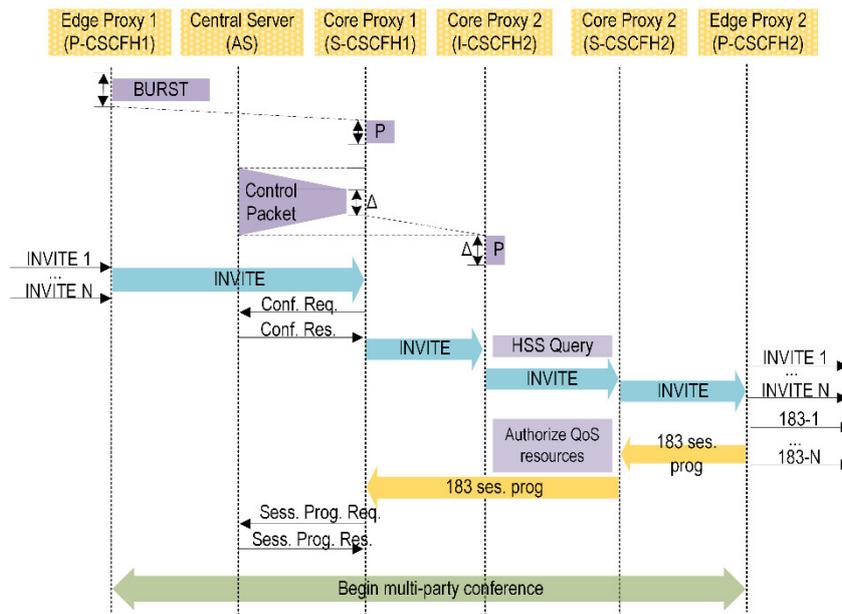
- UE (User Equipment), which takes the role for a request/response pair representing end users.
- central server known as application server (AS),
- SIP proxy servers (CSCFs) which includes edge-proxy (P-CSCF) and core-proxy (I-CSCF and S-CSCF).

Call flow diagram for conferencing session establishment is show in Fig. 1. Let us consider the procedure with more details. A group of UEs initiate a number of requests to try to establish a conferencing service and send INVITE messages towards Edge Proxy 1. All these messages should be delivered to the same P-CSCF server and can be aggregated by Edge Proxy 1 and send towards the target server as a single INVITE request with the list of participants.

When the INVITE message reaches Edge Proxy 2, the server will generate the required number of INVITE requests and send them towards the target UEs. Many 183 Session in Progress responses reach Edge Proxy 2 mean session establishment in progress. When the last response reaches Edge Proxy 2, the server will send a single 183 Session in Progress response towards the Edge Proxy 1.

The implementation of grouping approach on Edge Proxies enhances the bandwidth the whole system but requires to install in the network Edge Servers with higher performance.

A packet can be blocked either at an Edge Proxy or Core Proxy. The Edge Proxy model is used to quantify the probability that a packet is blocked at an Edge Server’s buffer. In case of multi-party conferencing, the requests arrive as a batch and the destinations are in the group therefore the Edge Proxy modeled as bulk queue with an exhaustive service with multiple vacations can be used to provide efficient packet aggregation.



Figures 1: Conferencing flow session setup

IMS CONFERENCING SERVER MODEL

Let us assume that customers arrive at a single-server queue with finite capacity buffer R and receive service in accordance with FCFS policy. The customers reach the system in a batch according the Poisson process with rate λ . The bath size is a random variable ξ with the following distribution function

$$P\{\xi = i\} = q_i, i = 1, \dots, K.$$

Let $\lambda_i = q_i \lambda$ be the intensity of arriving a batch with i - customers. If there is enough space in the buffer customers stand in a queue, otherwise they will be dropped.

Let n denote the buffer occupancy. Customers are aggregated in the buffer and receive service. Due to server capacity limitations, it processes up to L customers from the buffer at once. Furthermore, server does not wait until the number of customers in the queue

exceeds L , but it processes $\min(L, n)$ customers. The processing time is exponentially distributed with the mean μ^{-1} , and it does not depend on the batch size.

The server takes working vacation at the time when the system is empty. During the vacation period new arrived customers are stored in the buffer. The server takes another new vacation if only there is no any new customer in the queue. The vacation time is a random variable which is exponentially distributed with parameter θ . While server is on the vacation, the buffer occupancy starts growing and if it exceeds the value H the vacation starts to process the customers from the buffer. Thus, the server operates in two modes: normal mode ($m = 0$) and vacations mode ($m = 1$), where m is the server operating status. The described system can be denoted as $M^{[X]} | M^{[Y]} | 1 | R | E, MV$.

The functioning of the system is described by the Markov process $\mathbf{X}(t) = (m(t), n(t))$ over the state space $\mathcal{X} = \{(m, n) : m \in \{0, 1\}, 0 \leq n \leq R\}$.

Clearly $\mathbf{X}(t)$ is ergodic and thus stationary distribution exists. Let $p_{m,n} = \lim_{t \rightarrow \infty} P(m(t) = m, n(t) = n)$ be the stationary distribution of the process. The non-diagonal and nonzero elements of infinitesimal operator $A = (a_{m,n})_{\substack{m=0,1 \\ n=0,\dots,R}}$ of the process $\mathbf{X}(t)$ can be written in the following way

$$a_{m,n} = \begin{cases} q_i \lambda, (m'', n'') = (m', n') + (0, i), \\ \left(\begin{matrix} (n' \leq R - i) \wedge \\ (i \leq K) \wedge \\ (m'' = m' = 0) \end{matrix} \right) \vee \left(\begin{matrix} (n' \leq R - i) \wedge \\ (i \leq K) \wedge \\ (m'' = m' = 1) \end{matrix} \right); \\ \mu, (m'', n'') = (m', n') - (\min(L, n'), 0), \\ (k' + k'' = 0) \wedge (n' > 0), \\ (m'', n'') = (0, 0), (m', n') = (1, 0); \\ \theta, (m'', n'') = (m', n') - (0, 1), (n' = n'') \wedge \\ (m' = 1) \wedge (m'' = 0). \end{cases}$$

Performance measures

Let π denote the blocking probability of batch

$$\pi = \frac{1}{\lambda} \sum_{i=0}^{K-1} \sum_{k=i+1}^K \lambda_k (p_{0,R-i} + p_{1,R-i}).$$

The utilization of the server $UTIL$ is given by the following formula

$$UTIL = \frac{1}{\mu} \sum_{k=1}^K k \lambda_k \left(1 - \sum_{i=R+1-k}^R p_{0,i} + p_{1,i} \right)$$

The mean number of customers in the queue can be calculated as

$$Q = \sum_{k=1}^R k (p_{0,k} + p_{1,k}).$$

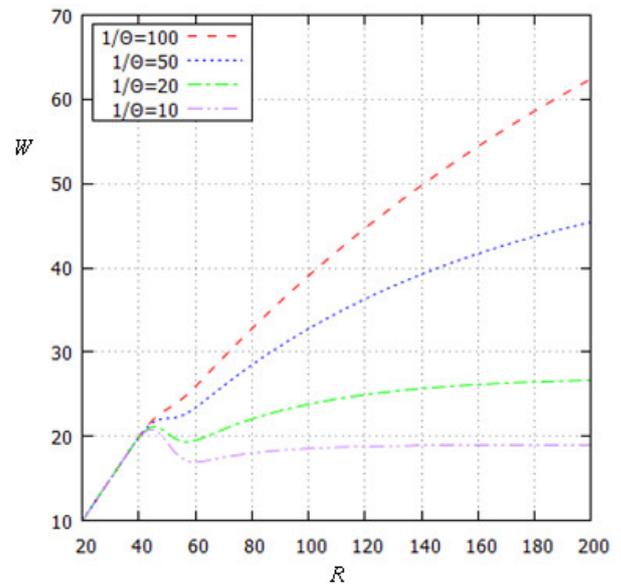
Using Little's formula, the mean waiting time in the queue is expressed as

$$W = \frac{Q}{\sum_{k=1}^K k \lambda_k \left(1 - \sum_{i=R+1-k}^R (p_{0,i} + p_{1,i}) \right)}$$

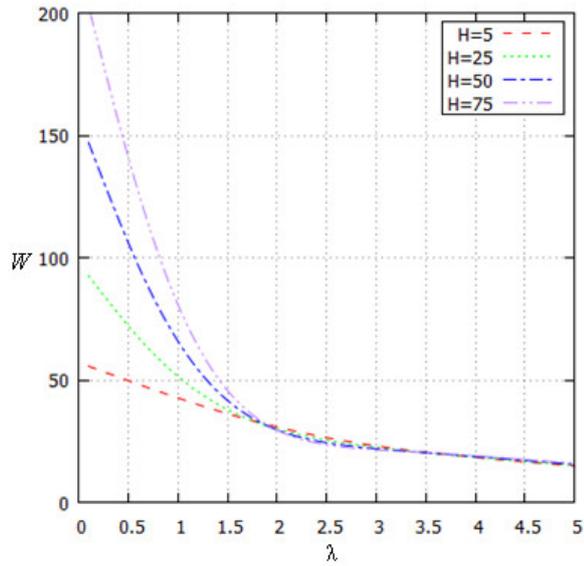
Numerical example

We perform several numerical experiments for a combination of different input parameters: $\lambda \leq 5$, θ^{-1} from 10 to 100, $\mu = 6$, $R = 100$, $K = 10$, $L = 5$.

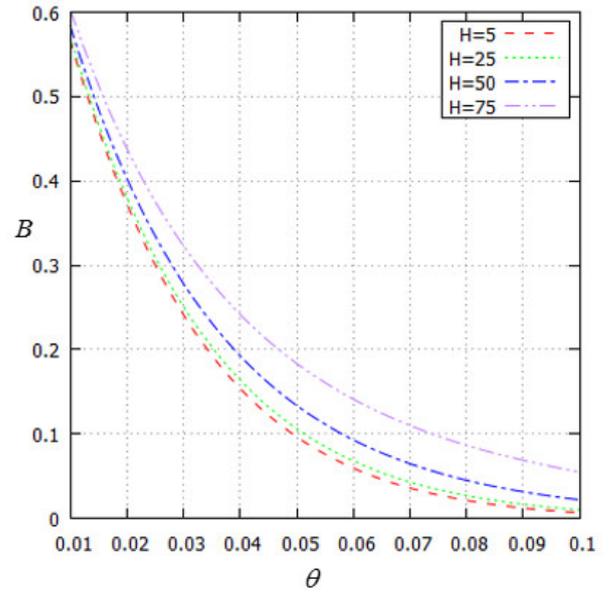
As shown in Fig. 2 as buffer size increases the mean waiting time increase from $R = 40$, but according Fig. 7 blocking probability asymptotically decrease. In observing Fig. 3 we see that increase of intensity rate λ does not make significant influence on mean waiting time for various value of threshold.



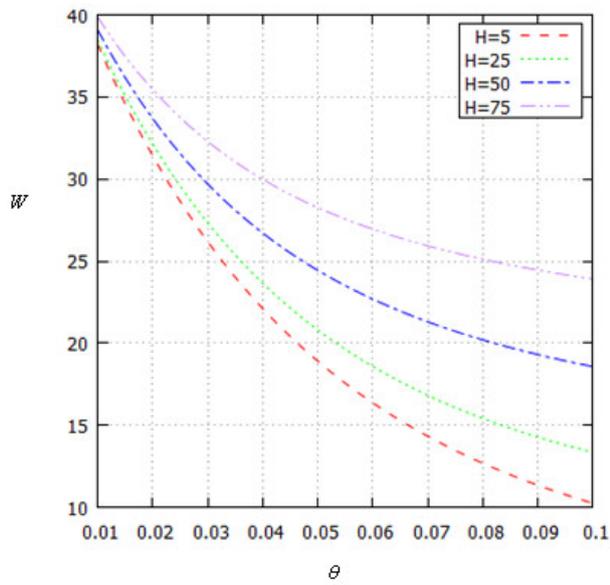
Figures 2: Mean waiting time on varying buffer size for different value of mean vacation time



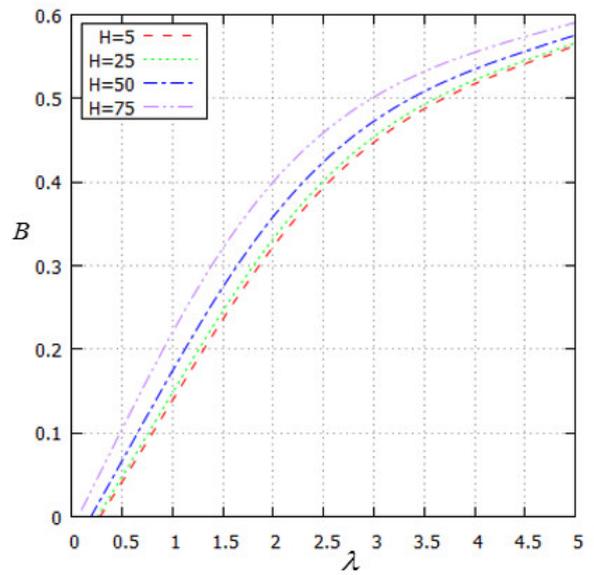
Figures 3: Mean waiting time on varying arrival rate for different value of the threshold



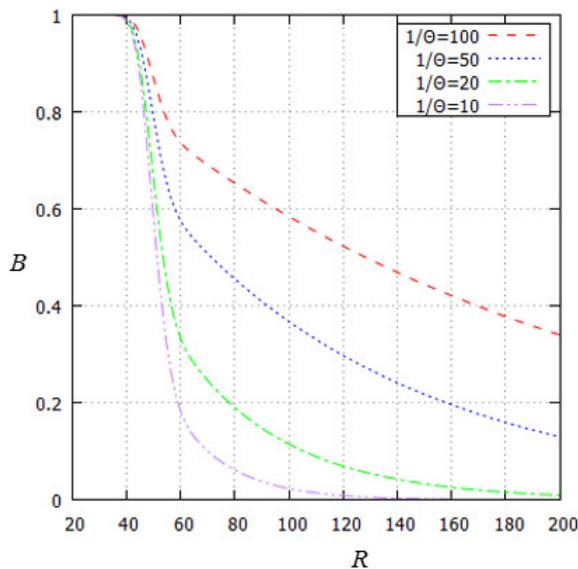
Figures 5: Blocking probability on varying mean vacation time for different value of the threshold



Figures 4: Mean waiting time on varying mean vacation time for different value of threshold



Figures 6: Blocking probability on varying arrival rate for different value of the threshold



Figures 7: Blocking probability on varying buffer size for different value of mean vacation time

SUMMARY AND FUTURE STUDY

In this paper, we give an overview of the IMS conferencing service architecture and a basic call flow. The queuing model with batch arrival batch exhaustive service and working vacations was analyzed.

The effectiveness of the server depends on external factors such as arrival traffic rate and on server policy configuration (value of θ , H) as well. For further study we propose the following optimization problem that will allow to increase IMS server performance

$$F = \begin{cases} \frac{\lambda(1-B(\theta, H))}{L\mu} \rightarrow \max, \\ B(\theta, H) \leq 10^{-3}, \\ W(\theta, H) \leq 10. \end{cases}$$

Notes and Comments. This work was supported in part by the Russian Foundation for Basic Research (grant 15-07-03608).

REFERENCES

- Camarillo, G., Garcia-Martin, M. A. 2008. The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds, 3rd ed. Wiley.
- Rosenberg, J., Schulzrinne, H., Camarillo, G. et al. 2002. SIP: Session Initiation Protocol. RFC 3261.
- IP Multimedia Subsystem (IMS), 2004. 3GPP, Stage 2 (Release 5). 3GPP TS 23.228, vol. 5.
- IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP). 2015. Stage 3 (Release 13), 3GPP TS 24.229.
- Tien, A. Le, Nguyen H. 2010. Centralized and distributed architectures of scalable video conferencing services. In the

Second International Conference on Ubiquitous and Future Networks, Jeju Island, Korea, pp. 394-399.

Pavel O. Abaev, Yuliya V. Gaidamaka, Konstantin E. Samouylov, Sergey Ya. Shorgin. 2013. Design and Software Architecture of SIP Server for Overload Control Simulation. Proceedings of the 27th European Conference on Modelling and Simulation, ECMS 2013, pp. 580-586.

Pavel Abaev, Yuliya Gaidamaka, and Konstantin E. Samouylov. 2012. Modeling of Hysteretic Signaling Load Control in Next Generation Networks. Lecture Notes in Computer Science. Germany, Heidelberg, Springer-Verlag, 2012, vol. 7469, pp. 440-452.

Mishra G., Dharmaraja S., Kar S. 2014. Performance Analysis of Multi-party Conferencing in IMS using Vacation Queues. IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1-6.

Rosenberg, J. 2008. Requirements for Management of Overload in the Session Initiation Protocol. RFC 5390.

AUTHOR BIOGRAPHIES

PAVEL ABAEV received his Ph.D. in Computer Science from the Peoples' Friendship University of Russia in 2011. He is an Assistant Professor in the Department of Applied Probability and Informatics at Peoples' Friendship University of Russia since 2013. His current research focus is on SDN/NFV, performance analysis of wireless 4G/5G networks and M2M communications, applied probability and queuing theory, and mathematical modeling of communication networks. His email address is: pabaev@sci.pfu.edu.ru.

VITALY BESCHASTNY received a BSc. degree in Applied Mathematics and Informatics in 2014 from People's Friendship University, Moscow, Russia. Currently he enrolled in MSc program of the Department of Applied Probability and Informatics at the same university. His present research focuses on performance analysis of SDN/NFV, resource management in D2D-enabled cellular networks, mathematical modeling and performance analysis of computer and communication systems. His email address is: vbeschastny@sci.pfu.edu.ru.

ALEXEY TSAREV is currently enrolled in BSc. program of Department of Applied Mathematics and Informatics at People's Friendship University, Moscow, Russia. His present research focuses on performance analysis of 5G networks, resource management in D2D-enabled cellular networks, mathematical modeling and performance analysis of computer and communication systems. His email address is: atsarev@sci.pfu.edu.ru.

NEW SCHEDULING POLICY FOR ESTIMATION OF STATIONARY PERFORMANCE CHARACTERISTICS IN SINGLE SERVER QUEUES WITH INACCURATE JOB SIZE INFORMATION

Lusine Meykhanadzhyan
Peoples' Friendship University of Russia,
Moscow, Russia
Email: lameykhanadzhyan@gmail.com

Rostislav Razumchik
Institute of Informatics Problems
of the FRC CSC RAS, Moscow, Russia,
Peoples' Friendship University of Russia,
Moscow, Russia
Email: rrazumchik@ipiran.ru

KEYWORDS

inaccurate job size, long tails, service policy, size-based scheduling

ABSTRACT

The study of size-based and size-oblivious scheduling policies with inaccurate job size information appears nowadays to be an important direction of scientific studies because as recent research results show advantages of size-based policies can be saved even when the job sizes are not perfectly known a priori. This paper is focused on the same topic but touches upon a different question: is it possible to predict such estimates of system's performance characteristics (for example, job's mean sojourn time), that will be close to those which will be observed in practice, if the scheduler is provided only with the inaccurate information about the job size distribution? It is shown here that there are conditions under which the answer to the question is positive. A simple mathematical model ($M/G/1$ queueing system) of a top level view of a data-intensive execution engine is being proposed. It is shown that, in case of long-tailed service time distribution, a special service policy – Preemptive-Last-Come-First-Served with service time re-generation on arrival instants – allows one to obtain better upper bounds for job's mean sojourn time than those achieved by common work conserving policies. Extensive numerical examples are presented.

INTRODUCTION

The topic of this paper concerns performance evaluation of data-intensive systems in the presence of job size estimation errors. It is known that celebrated size-based scheduling policies which allow one to increase system's performance by changing the service policy start to perform poor (compared to size-oblivious policies like Processor sharing) as the uncertainty about the job size distribution increases (see Lu et al. (2004)). Since the inaccuracy in job size information is reported to be quite a common issue in practice (see Dell'Amico (2013); Dell'Amico et al. (2015); Wierman (2008)) there appeared a number of research papers appeared that fo-

cused on the behaviour of size-based scheduling policies with inaccurate size estimations. To our knowledge one can find in Dell'Amico et al. (2015); Wierman (2008); Harchol-Balter et al. (2003); Chang et al. (2011) the most recent results on this topic (including short reviews).

An important question which is being faced in systems with inaccurate job size information is whether it is possible to devise a scheduler which performs as good as the scheduler fed with the accurate information. A number of studies show, that sometimes it is possible (Dell'Amico et al. (2015)). In this paper we try to look at the performance of such systems from a different point of view and raise another question. We start with an example. Assume that one has a simulation or a physical copy of the data-intensive system. Each job is admitted into the system and after being served departs from it, and never comes back. In practice this system will be fed for a long time with the flow of jobs of random size s , exact distribution of which is unknown to the system but is known to be long-tailed. Denote the mean sojourn time which will be seen in the future by ν^* . Before launching the system the owner is interested in job's mean sojourn time which can be guaranteed by the system. In order to find it out one can run simulation experiments with one (or more) scheduler and some job's traces which contain *user's estimated* job's sizes which we denote by \hat{s} . Denote by ν the estimated mean sojourn time. Clearly ν is only the estimate for the unknown value ν^* . The questions are: (i) is it possible to improve the estimate ν by using only the available information about the distribution of \hat{s} and having control over the scheduler? (ii) if it is possible then what is the quality of the estimate? In this paper one shows that in some cases (when the distribution of \hat{s} is long-tailed) the answer to the question (i) is positive. The quality of estimate is investigated numerically.

As the answer to the question (i) may depend on many technical aspects related to real-life systems we will restrict ourselves to a probably the most simple case: a system is modelled by a single-server queue of $M/G/1$ type. An extent to which such assumption is an oversimplification of the real-life systems can be seen, for example, from papers Lu et al. (2004); Qiao et al. (2004); Dell'Amico et al. (2015) where it was used to evaluate

size-based scheduling policies with inaccurate job size information. Coming back to the question (i), we suggest to use a special service policy which utilizes the assumption¹ that job's true service time distribution is long-tailed. The idea of the policy is the following. Suppose that upon each arrival of a job the processing of current running job is interrupted and one re-generates the service time of both customers (depending on their current service times and according to the distribution of δ) and then the service is resumed with the new service time and the arrived customer occupies one place in the queue. A number of questions arise here. Will such service policy lead to a better estimate of v , than the ordinary work conserving policy like processor sharing? If yes then under what conditions? One can expect this policy to be efficient sometimes because long-tailed distributions "tend" to have decreasing hazard rates and increasing mean residual lives. As the numerical experiments show this is really so.

In this paper we propose a simple mathematical model, which is a top level view of a data-intensive execution engine and thus does not take into consideration most of its technical details. Specifically one considers an execution engine as an $M/G/1$ queueing system, with an infinite queue, a single flow of jobs, i.i.d. execution (service) times and pre-emptive last come first served discipline which allows service time re-generation on arrival instants. The comments to assumptions allowing one to take such a simplified view can be found in papers Dell'Amico (2013); Dell'Amico et al. (2015), devoted to the design of the new scheduler for Hadoop execution engine for data-intensive systems. At first we present some analytical results concerning the analysis of system's stationary characteristics, concentrating on sojourn time distribution. These results heavily rely on Meykhanadzhyan et al. (2014) and thus are presented in short. Then we show that if the service time distribution is long-tailed (either s or δ) PLCFS-re policy allows one to obtain upper bounds for the true value v^* , which can be much better than the values obtained using common work conserving disciplines. The comparison is done with classic size-based and size-oblivious service policies, which use the inaccurate job size information for scheduling.

In the next section the detailed description of the mathematical model is presented, which is followed by some results concerning the system's stationary performance characteristics. Section 3 is devoted to numerical results. In the conclusion one discusses in short the obtained results and directions of further research.

MATHEMATICAL MODEL

Consideration is given to a queueing system with one queue on infinite capacity, one server and Poisson arrival flow of rate λ . Service times of customers are i.i.d. ran-

¹The practical evidences for such an assumption can be found, for example, in Ren et al. (2012); Crovella (2001).

dom variables with known cumulative distribution function $B(x)$ and density $b(x) = B'(x)$. We assume that the customer's service time becomes known upon its arrival at the system and at any time instant the service time (remaining service time) of each customer in the system (both in the server and in the queue) is known. Newly arriving customers and those which are in the queue obey the special service policy which we will call Preemptive-Last-Come-First-Served with re-service (PLCFS-Re). It implies the following service rule. When a customer arrives at the system it interrupts the service process of the customer in server (if any) and compares its service time u with the (remaining) service time v of the customer in server. Then the arrived customer receives new service time U and the customer in service receives new service time V according to the known distribution

$$D(x, y|u, v) = \mathbf{P}\{U < x, V < y|u, v\}$$

which can depend on u and v . After that the service of the customer in server is resumed with the updated service time V and the arrived customer updates its service time with U and occupies the last place in the queue. When the service time of a customer in service becomes zero it leaves the system and one customer from the last place in the queue enters server. For the sake of convenience it is assumed that the density $d(x, y|u, v) = \partial^2 D(x, y|u, v)/(\partial x \partial y)$ is continuous and bounded. Note that for each u and v the following identity holds:

$$\int_0^\infty \int_0^\infty d(x, y|u, v) dx dy = D(\infty, \infty|u, v) = 1. \quad (1)$$

Let the stationary regime of the system exist. The main performance characteristic under study in this paper is the system's stationary mean sojourn time. Denote by $\beta(s)$ the Laplace-Stieltjes transform (LST) of $B(x)$ i.e.

$$\beta(s) = \int_0^\infty e^{-sx} dB(x).$$

For example, if $d(x, y|u, v) = b(x)b(y)$ i.e. the new service times of the arriving customer and the customer in server are chosen (upon arrival) independently of their previous service times u and v the stationary regime (and the mean sojourn time) exists if and only if the inequalities $1/2 < \beta(\lambda) < 1$ hold.

Stationary probabilities

Denote by $\nu(t)$ the number of customers in the system at instant t , and by $\vec{\xi}(t) = (\xi_1(t), \dots, \xi_{\nu(t)}(t))$ — the row vector, in which $\xi_1(t)$ is the (remaining) service time of the customer in server, $\xi_2(t)$ — the service time of the 1st customer in the queue, \dots , $\xi_{\nu(t)-1}(t)$ — the service time of the last, $(\nu(t) - 1)$, customer in the queue. If $\nu(t) = 0$ the vector $\vec{\xi}(t)$ is not defined. Then the process $\eta(t) = (\nu(t), \vec{\xi}(t))$ describing the evolution of the number of customers in the system is a continuous-time Markov chain.

Let us introduce the stationary distribution of the chain $\eta(t)$:

$$p_0 = \lim_{t \rightarrow \infty} \mathbf{P}\{v(t) = 0\},$$

$$P_n(x_1, \dots, x_n) = \lim_{t \rightarrow \infty} \mathbf{P}\{v(t) = n, \xi_1(t) < x_1, \dots, \xi_n(t) < x_n\}, \quad n \geq 1.$$

From the system's description and introduced assumptions it can be shown that density functions

$$p_n(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} P_n(x_1, \dots, x_n), \quad n \geq 1,$$

are continuous and bounded. Using the properties of the restricted Markov chains and the properties of the PLCFS-Re service policy one can write out the system of integro-differential equations for the stationary density functions $p_n(x_1, \dots, x_n)$. The details of this approach can be found in Meykhanadzhyan et al. (2014). But in order to compute the mean sojourn time it is enough to find the marginal density functions $p_1(x)$ and

$$p_n(x) = \int \dots \int_{x_2, \dots, x_n > 0} p_n(x, x_2, \dots, x_n) dx_2 \dots dx_n, \quad n \geq 2,$$

which take into consideration only the number of customers in the system and the remaining service time of the customer in server. Referring again to the approach in Meykhanadzhyan et al. (2014), one can obtain the following system of integro-differential equations for the functions $p_n(x)$:

$$-p'_n(x) = a_n(x) - \lambda p_n(x) + \int_0^\infty K_n(x, v) p_n(v) dv, \quad n \geq 1, \quad (2)$$

where $a_1(x) = \lambda b(x)p_0$,

$$a_n(x) = \lambda \int_0^\infty p_{n-1}(v) dv \int_0^\infty b(u) du \int_0^\infty d(y, x|u, v) dy, \quad n \geq 2,$$

$$K_n(x, v) = \lambda \int_0^\infty b(u) du \int_0^\infty d(x, y|u, v) dy, \quad n \geq 1.$$

The initial conditions for the system (2) follow from the properties of density functions and have the form

$$\lim_{x \rightarrow \infty} p_n(x) = 0, \quad n \geq 1.$$

Note that for an arbitrary function $d(x, y|u, v)$ the system (2) can be solved numerically. Substitution of $p_n(x) = e^{\lambda x} q_n(x)$ into (2) and subsequent integration of the new system for $q_n(x)$ lead to Fredholm equations of the second kind with non-negative kernels, for which standard approaches are still feasible (for example, iteration with first iteration equal to zero). The only unknown probability left is the probability of the empty system p_0 , which can be computed from the normalization condition

$$\sum_{n=0}^{\infty} p_n = 1,$$

where

$$p_n = \int_0^\infty p_n(x) dx, \quad n \geq 1,$$

is the stationary probability of n customers in the system.

In some cases, which are relevant for practice (see the numerical section), one can easily compute some useful quantities from (2). Consider the following special case of the PLCFS-Re service policy. Let upon arrival of a new customer its service time u and the (remaining) service time v of the customer in server (if any) be regenerated independently of u and v (i.e. independently of how long the customer has been already served) according to $B(x)$ (i.e. $d(x, y|u, v) = b(x)b(y)$). Then the system (2) can be simplified to the following form:

$$-p'_n(x) = \lambda b(x)p_{n-1} - \lambda p_n(x) + \lambda b(x)p_n, \quad n \geq 1. \quad (3)$$

Remarks to the solution of this system remain the same as to the system (2). But here one is able to find the explicit expression for the mean number of customers in the system without finding all the unknown functions $p_n(x)$. Indeed, if one introduces the generating function

$$\pi(z, x) = \sum_{n=1}^{\infty} p_n(x) z^n,$$

then, if one treats z as a parameter, the solution of the system (3) in terms of $\pi(z, x)$ can be written in the form

$$\pi(z, x) = \int_x^\infty e^{-\lambda(u-x)} [\lambda z b(u) p_0 + \lambda(1+z)b(u)\pi(z)] du, \quad (4)$$

where $\pi(z) = \int_0^\infty \pi(z, v) dv$. By integrating (4) from zero to infinity and putting $z = 1$, one obtains the equation for the determination of $\pi(1)$. Its solution is $\pi(1) = [\beta(\lambda)]^{-1} - 1$ and thus $p_0 = 1 - \pi(1) = 2 - [\beta(\lambda)]^{-1}$. Next, by differentiating (4) and then integrating out x , one obtains the equation for the mean number of customers in the system N , which solution is $N = [1 - \beta(\lambda)]/[2\beta(\lambda) - 1]$.

Stationary sojourn time

As one is unable to use the Little's law without the prior check for the considered system, we now dwell on the stationary sojourn time distribution. Denote by $\chi(s)$ the LST of the customer's sojourn time and by $u(s)$ the LST of the system's busy period.

For the sake of simplicity we will consider only the special case of the PLCFS-Re service policy when $d(x, y|u, v) = b(x)b(y)$, because this allows one to obtain most of the expressions in explicit form. The general case of $d(x, y|u, v)$ can be handled in a similar manner, yet the LST $\chi(s)$ and $u(s)$ can be obtained only as solutions of certain functional equations.

Following again the approach from Meykhanadzhyan et al. (2014) and using the properties of the PLCFS-Re service policy one can verify that the LST of the busy period $u(s)$ satisfies the equation

$$u(s) = \beta(s + \lambda) + \frac{\lambda}{\lambda + s} [1 - \beta(s + \lambda)] u^2(s),$$

and its appropriate solution has the form

$$u(s) = \frac{\lambda + s - \sqrt{[\lambda + s]^2 - 4\lambda[1 - \beta(s + \lambda)]\beta(s + \lambda)[\lambda + s]}}{2\lambda[1 - \beta(s + \lambda)]}.$$

It can be verified that the busy period is finite with probability 1 (i.e. $u(0) = 1$) if and only if $\beta(\lambda) > 1/2$ and the mean length of the busy period $-\beta'(0)$ is finite if and only if $1/2 < \beta(\lambda) < 1$.

Denote by $\psi(s)$ the LST of the total time which the customer spends in service once it enters server (this includes all possible re-regenerations of its service time). Using the first step analysis it can be shown that $\psi(s)$ satisfies the equation

$$\psi(s) = \beta(\lambda + s) + \frac{\lambda}{\lambda + s} \psi(s)[1 - \beta(\lambda + s)],$$

and thus has the form

$$\psi(s) = \frac{(\lambda + s)\beta(\lambda + s)}{s + \lambda\beta(\lambda + s)}.$$

Remembering that customers from the queue are served according to LCFS order then, using the law of total probability, one obtains the following expression for the LST $\chi(s)$ of the customer's sojourn time distribution:

$$\chi(s) = p_0\psi(s) + (1 - p_0)\psi(s)u(s). \quad (5)$$

Further computation of the mean and higher moments (if they exist) of the stationary sojourn time is straightforward.

NUMERICAL RESULTS

In this section we demonstrate through numerical examples that in certain cases PLCFS-Re service policy is able to eliminate job-size estimation errors and produce good upper bounds for values of the mean sojourn time that are seen in the system with true job-sizes.

In order to be able to make comparisons we assumed that the true job sizes (service times), say S , have a Weibull distribution with the shape parameter k and the scale parameter α . Notice that such assumption is justified by the empirical measurements (see, for example, Dell'Amico et al. (2015) and references therein). Here we present results both for long-tailed and light-tailed service time distributions achieved by varying only the shape parameter k . The values of the scale parameter α were always set to guarantee that the true mean service time $E(S)$ is equal to 1.

In the experiments we used the assumption, which has already been adopted in a number of recent papers on the topic (see Dell'Amico (2013); Dell'Amico et al. (2015)), when the job-size estimation error, say X , is log-normally distributed with zero mean and variance σ^2 and a job, having true size S , is estimated as $\hat{S} = SX$. Denote by $B(x)$ the cumulative distribution of the random variable \hat{S} . The mean $E(\hat{S})$ of the distribution is always greater than 1 and as σ grows the sizes of the jobs are biased towards overestimating.

To summarize, one considers $M/G/1$ infinite capacity queueing system with Poisson arrivals of rate λ and i.i.d. service times with distribution $B(x)$. Assume that four different service policies are implemented in the system: Processor Sharing (PS), Shortest-Preemptive-Last-Come-First-Served (SPLCFS), First-Come-First-Served (FCFS) and PLCFS-Re. Denote by v^{PS} , v^{SPLCFS} , v^{FCFS} , $v^{\text{PLCFS-Re}}$ the values of customer's stationary mean sojourn time under each of the four policies. We are interested in two aspects:

- understanding the relationships between $v^{\text{PLCFS-Re}}$ and v^{PS} , v^{SPLCFS} , v^{FCFS} in the presence of estimation errors i.e. when $\sigma > 0$;
- understanding how close the values of $v^{\text{PLCFS-Re}}$ are to the values of true stationary mean sojourn time, which we denote by² $v^{*\text{PS}}$, $v^{*\text{SPLCFS}}$ and $v^{*\text{FCFS}}$.

As the considered model is of $M/G/1$ type, for the computation of v^{PS} , v^{SPLCFS} and v^{FCFS} one can use known analytic expressions.

According to the specification of the PLCFS-Re service policy (see Section 2), in order to use it one has to specify the function $d(x, y|u, v)$. The flexibility of the definition of $d(x, y|u, v)$ gives a plenty of choices but we will choose probably the most simple one. Let $d(x, y|u, v) = b(x)b(y)$ i.e. the service time of the newly arriving customer and the (remaining) service time of the customer in server (if any) are re-generated independently of their current values. Then the value of $v^{\text{PLCFS-Re}}$ is equal to $-\chi'(0)$, which can be computed using (5). Notice that being defined in such a way, this policy is completely memoryless and not work conserving.

In Fig. 1, Fig. 2, Fig. 3 and Fig. 4 one can see the relationship between the values of $v^{\text{PLCFS-Re}}$, v^{PS} and true mean sojourn time $v^{*\text{PS}}$. Along the x-axis one shows the values of load $\lambda E(S) = \lambda$ in the system without estimation errors. The values of mean sojourn time are given along the y-axis.

From Fig. 1 one can see that if the shape parameter is $k = 1$ (i.e. the true service time distribution is exponential) then $v^{*\text{PS}} < v^{\text{PLCFS-Re}} < v^{\text{PS}}$ over the whole range of load values. In the presence of errors $v^{\text{PLCFS-Re}}$ provides a good upper bound for the true mean sojourn time $v^{*\text{PS}}$ under PS policy. Another interesting observation from Fig. 1 is the following. If one fixes the arrival rate, say $\lambda = 0.9$, then starting from a certain value of σ , one is already unable to compute v^{PS} because system's load exceeds 1 (in case $\sigma = 1$ and $\lambda = 0.9$ system's load is ≈ 1.48). But the value $v^{\text{PLCFS-Re}}$ still can be computed and provides a good upper bound for the true value $v^{*\text{PS}}$. Thus sometimes PLCFS-Re policy allows one to make estimations of the true values of mean sojourn time for wider ranges of system's load. Almost similar dynamics can be observed for higher values of k (see Fig. 2).

²These denote the values of customer's stationary mean sojourn time in case of no estimation errors i.e. when the service time distribution is Weibull.

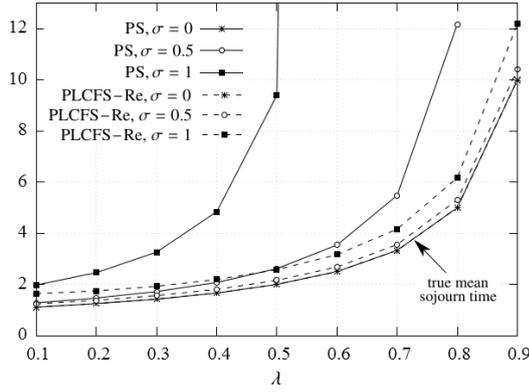


Figure 1: Mean sojourn time versus load for PS and PLCFS-Re service policies. True service time distribution is exponential ($k = 1$). The inequalities $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ always hold.

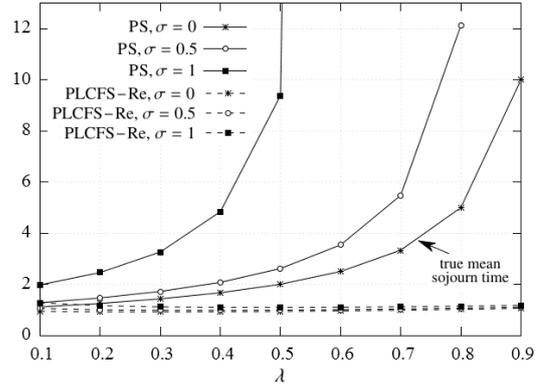


Figure 4: Mean sojourn time versus load for PS and PLCFS-Re service policies. True service time distribution is long-tailed ($k = 0.5$) with joint ratio $\approx 20/80$. PLCFS-Re with $d(x, y|u, v) = b(x)b(y)$ almost always underestimates the true values of the mean sojourn time.

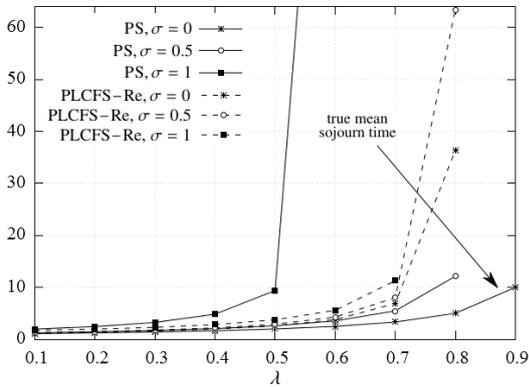


Figure 2: Mean sojourn time versus load for PS and PLCFS-Re service policies. True service time distribution is light-tailed ($k = 1.5$). For high values of estimation errors $v^{*PS} < v^{PLCFS-Re} < v^{PS}$.

ity $v^{PLCFS-Re} < v^{PS}$ almost always holds. But now the range of load values for which $v^{PLCFS-Re}$ provides an upper bound for the true value v^{*PS} depends on the estimation error σ : the greater σ the wider range. Yet for quite low values of k (see Fig. 4) the values $v^{PLCFS-Re}$ significantly underestimate mean sojourn time for PS service policy and are thus useless.

Qualitatively the above observations are valid for FCFS and SPLCFS policies as well (see Fig. 5 for SPLCFS policy).

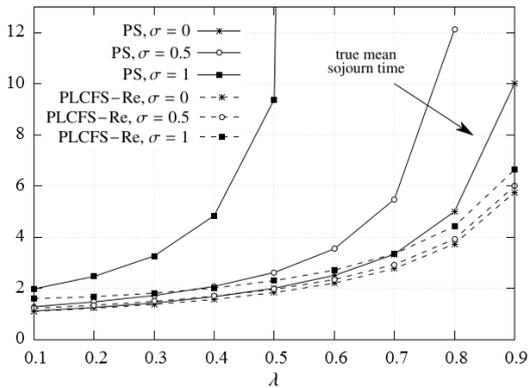


Figure 3: Mean sojourn time versus load for PS and PLCFS-Re service policies. True service time distribution is long-tailed ($k = 0.9$) with joint ratio $\approx 30/70$. The values of load for which $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ depend on the estimation error σ .

As the tail of the true service time distribution becomes heavier the situation changes. From Fig. 3 one can see that for $k = 0.9$ in the presence of errors the inequal-

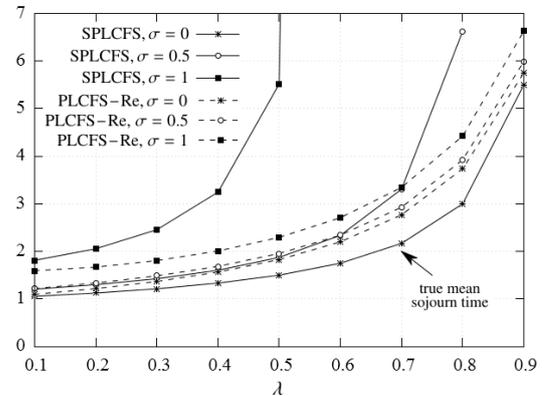


Figure 5: Mean sojourn time versus load for SPLCFS and PLCFS-Re service policies. True service time distribution is long-tailed ($k = 0.9$) with joint ratio $\approx 30/70$. PLCFS-Re with $d(x, y|u, v) = b(x)b(y)$ for high values of load provides better estimates of v than ordinary SPLCFS.

From Fig. 1–4 one can observe that for some values of k and σ the inequalities $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ are held and for others are not. More thorough study has revealed that these inequalities also depend on the value of λ (see Fig. 6).

According to our experiments the PLCFS-Re policy can provide good upper bounds for the true values of the mean sojourn time only if the service time distribu-

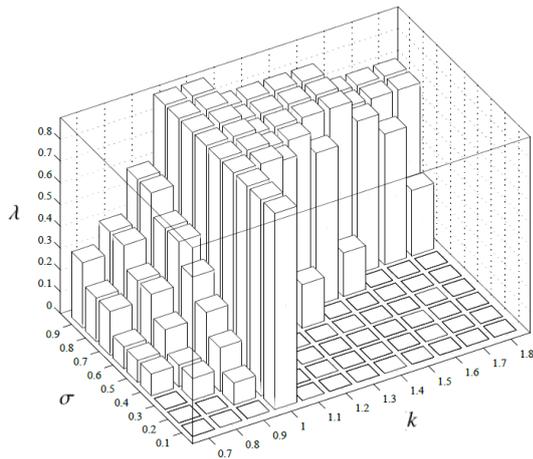


Figure 6: A visual presentation of the value range of λ , σ and k within which $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ i.e. when PLCFS-Re policy always outperforms PS. The presence of the vertical bar and its height indicates the set of values of λ , σ and k for which $v^{*PS} < v^{PLCFS-Re} < v^{PS}$. The absence of the bar for the values of λ , σ and k indicates that PLCFS-Re policy is worse than PS. For the ease of presentation it is assumed that λ , σ and k take only discrete values (with step 0.1). We note that not the full range of possible values of λ , σ and k is displayed.

tion $B(x)$ is long-tailed and thus exhibits more or less the mass-count disparity³. If one chooses the joint ratio and median-to-median distance⁴ as a characterisation of the service time distribution $B(x)$ then from Table 1 it can be seen that PLCFS-Re policy is useful (i.e. guarantees that $v^{*PS} < v^{PLCFS-Re} < v^{PS}$) whenever the joint ratio is approximately less than 32. But even for joint ratio < 32 the PLCFS-Re policy can underestimate the values of the true mean sojourn time, because there is also a dependency on the values of λ (see Fig. 6).

As the mass-count disparity becomes stronger (i.e. the tail of the service time distribution becomes more pronounced) PLCFS-Re policy with $D(x, y|u, v) = B(x)B(y)$ provides very bad (highly underestimated) values of the true mean sojourn time for all three policies PS, SPLCFS and FCFS. But simple changes in the definition of the function $D(x, y|u, v)$ allow one to broaden the the range of values k , σ and λ for which PLCFS-Re policy is good. For example, fix $M > 0$ and let the $d(x, y|u, v)$ be equal to

$$d(x, y|u, v) = \begin{cases} b(x)\delta(y - v), & u > M, v < M, \\ b(y)\delta(x - u), & u < M, v > M, \\ \delta(x - u)\delta(y - v), & u < M, v < M, \\ b(x)b(y), & u > M, v > M, \end{cases} \quad (6)$$

³A small number of samples account for the majority of mass, whereas all small samples together only account for negligible mass.

⁴Joint ratio is the generalization of the Pareto principle: $p\%$ of customers account for $(100-p)\%$ of the service time, and $(100-p)\%$ of service times account for $p\%$ of customers. Median-to-median distance is the result of the division of the median of the mass distribution by the median of count distribution. See, for example, (Feitelson, 2015, Chapter 5).

where $\delta(x)$ is the Dirac delta function. According to Fig. 6 PLCFS-Re policy with $d(x, y|u, v) = b(x)b(y)$ is worse than PS in the presence of errors when $k = 0.7$, $\sigma = 0.9$ and $\lambda = 0.4$. But if one uses the new definition (6) for $d(x, y|u, v)$ then it holds⁵ that $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ if $M = 244$.

CONCLUSION

The function $d(x, y|u, v)$, governing the PLCFS-Re service policy, presents a sort of control over the system. Indeed, the solution of the system (2) leads to Fedholm's equations of the second kind and each of them can be seen as a dynamic programming equation (Bellman's equation) but without the control variable. Numerical results show that even simple control without any memory (i.e. when $d(x, y|u, v) = b(x)b(y)$) allows one to eliminate errors in the estimations of the job sizes and obtain better upper bounds for the true mean sojourn time in $M/G/1$ type systems with inaccurate job size information. Here by saying "better" we mean "better than the values of the mean sojourn time that can be obtained using ordinary work conserving policies PS, LCFS, FCFS, PLCFS and some others". But such simple control does not work for too long-tailed service time distributions, exhibiting strong mass-count disparity (for example, exhibiting 20/80 Pareto principle). But by changing the definition of the function $d(x, y|u, v)$ (for example, by introducing a threshold) one can broaden the ranges of mass-count disparity in which the PLCFS-Re service policy outperforms ordinary policies. Up to now we were unable to find a general rule which specifies when the PLCFS-Re service policy can provide a good upper bound in the presence of errors for the true mean sojourn time. For each combination of system's initial parameters one has to compare the values manually. But in the experiments we have observed such properties as linear order (with respect to the policies) and monotonicity with respect to initial parameters. Even though the considered model can be considered as oversimplified, it allows one to reach the problem from an analytic point of view. In the presence of long-tailed distributions, which complicate the simulation experiments, this can be seen as an advantage. We see the following further directions of research: searching for the best form of the function $d(x, y|u, v)$; checking the feasibility of the proposed approach for other performance characteristics (such as variance of the mean sojourn time) and multi-server systems; evaluation of PLCFS-Re service policy when other approach for modelling of job's inaccurate size is used (as discussed in Tsafir et al. (2005)).

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grants 15-07-03007 and 15-07-03406).

⁵The exact values are $v^{*PS} = 1.66$, $v^{PLCFS-Re} = 1.73$, $v^{PS} = 3.75$.

Table 1: Values of the joint ratio and median-to-median distance for different values of k and σ . The scale parameter of the Weibull distribution is fixed and equal to $\alpha = \Gamma(1 + 1/k)^{-1}$ i.e. the mean of the Weibull distribution is 1. The values for which $v^{*PS} < v^{PLCFS-Re} < v^{PS}$ i.e. when PLCFS-Re policy outperforms PS are in bold type.

σ	$k = 0.7$		$k = 1$		$k = 1.8$	
	joint ratio $p\%/(1-p)\%$	m-m dist.	joint ratio $p\%/(1-p)\%$	m-m dist.	joint ratio $p\%/(1-p)\%$	m-m dist.
0	26.55/73.45 ^a	4.87	31.92/68.08 ^b	2.42	38.68/61.32 ^c	1.38
0.1	26.49/73.51	4.91	31.86/68.14	2.44	38.48/61.52	1.39
0.2	26.31/73.69	5.05	31.49/68.51	2.52	38.01/61.99	1.43
0.3	25.89/74.11	5.33	30.98/69.02	2.65	37.39/62.61	1.51
0.4	25.44/74.56	5.68	30.45/69.55	2.84	36.43/63.57	1.63
0.5	24.93/75.07	6.25	29.55/70.45	3.12	35.13/64.87	1.78
0.6	24.24/75.76	7.00	28.69/71.31	3.50	33.91/66.09	1.99
0.7	23.52/76.48	8.01	27.69/72.31	4.00	32.50/67.50	2.28
0.8	22.72/77.28	9.34	26.71/73.29	4.66	31.04/68.96	2.66
0.9	21.90/78.10	11.13	25.60/74.40	5.55	29.59/70.41	3.16

^aLong-tailed Weibull service time distribution with mean 1.

^bExponential service time distribution with mean 1.

^cLight-tailed service time distribution with mean 1.

REFERENCES

- Chang, H., Kodialam, M. S., Kompella, R. R., Lakshman, T. V., Lee, M., Mukherjee, S. 2011. Scheduling in MapReduce-like systems for fast completion time. In INFOCOM, Proceedings IEEE. New York. Pp.3074–3082.
- Crovella, M. 2001. Performance Evaluation with Heavy Tailed Distributions. JSSPP '01 Revised Papers from the 7th International Workshop on Job Scheduling Strategies for Parallel Processing. Pp.1–10.
- Dell'Amico, M. 2013. A simulator for data-intensive job scheduling. Tech. Rep. arXiv:1306.6023, arXiv.
- Dell'Amico, M., Carra, D., Michiardi, P. 2015. PSBS: Practical Size-Based Scheduling. IEEE Transactions on Computers. Issue.99. Pp. 1–15.
- Feitelson, D. 2015. Workload Modeling for Computer Systems Performance Evaluation. Cambridge University Press.
- Harchol-Balter, M., Schroeder, B., Bansal, N., Agrawal, M. 2003. Size-based scheduling to improve web performance. ACM Transactions on Computer Systems (TOCS). Vol. 21. Issue 2. Pp.207–233.
- Lu, D., Sheng, H., Dinda, P. 2004. Size-based scheduling policies with inaccurate scheduling information. Modeling, Analysis, and Simulation of Computer and Telecommunications Systems. (MASCOTS 2004). Proceedings. The IEEE Computer Society's 12th Annual International Symposium on. Pp.31–38.
- Meykhanadzhyan, L. A., Milovanova, T. A., Pechinkin, A. V., Razumchik, R. V. 2014. Stationary distribution in a queueing system with inverse service order and generalized probabilistic priority. Inform. Primen. Vol. 8. Issue 3. Pp.28–38. (in Russian)
- Qiao, Yi, Lu, D., Bustamante, F., Dinda, P. 2004. Looking at the server side of peer-to-peer systems. In Proceedings of the 7th workshop on Workshop on languages, compilers, and run-time support for scalable systems (LCR '04). ACM, New York, NY. USA. Pp.1–8.
- Ren K., Kwon, Y., Balazinska, M., Howe, B. 2012 Hadoops adolescence: A comparative workload analysis from three research clusters. In Technical Report, CMU-PDL-12-106.
- Tsafirir, D., Etsion, Y., Feitelson, D. G. 2005. Modeling user runtime estimates. In 11th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP), Springer-Verlag. Lect. Notes Comput. Sci. Vol. 3834. Pp. 1–35.
- Wierman, A., Nuyens, M. 2008. Scheduling despite inexact job-size information. In ACM SIGMETRICS Performance Evaluation Review. Vol. 36. Issue 1. Pp.25–36.

AUTHOR BIOGRAPHIES

LUSINE A. MEYKHANADZHYAN is a PhD student at the department of applied probability and informatics of Peoples Friendship University of Russia. At present her research interests are focused on queueing-theoretic aspects of scheduling. Her email address is lameyghanadzhyan@gmail.com.

ROSTISLAV V. RAZUMCHIK received his Ph.D. degree in Physics and Mathematics in 2011. Since then, he has worked as a senior research fellow at Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS). At present he holds the position of Head of the Information and Telecommunication System Modelling section at the FRC CSC RAS and associate professor position at Peoples' Friendship University of Russia. His current research activities are focused on queueing theory and its applications in performance evaluation of stochastic systems. His email address is rrazumchik@ipiran.ru.

AUTHOR INDEX

- 685, 692 Abaev, Pavel
705
614 Abusharekh, Ashraf M.
393 Adamek, Milan
114 Ambrosino, Daniela
107 Arantes, Silvia
80 Azab, Ahmed E.
159 Bagirova, Anna
407 Barbey, Hans-Peter
131 Barra, Carlos
692 Begishev, V.
669 Bening, Vladimir E.
333 Bentaleb, Toufik
474 Bernardon, Daniel P.
705 Beschastny, Vitaly
579 Bianchini, Devis
488 Bley, Andreas
145 Blueschke, Dmitri
279, 287 Bobal, Vladimir
300, 307
313, 347
388 Bonilla, Luis
534, 544 Bye, Robin T.
554, 564
495 Bzowski, Krzysztof
107 Candeias, Fatima
131 Canessa, Enrique
138 Cao, Vu Lam
453 Carletti, Luca
131 Chaigneau, Sergio E.
515 Chalardkid, Noppachai
293 Chalupa, Petr
59 Chandrasekaran, S.
635 Chodak, Grzegorz
508 Chotpan, Nattapong
628 Cicala, Giuseppe
18, 73 Cicirelli, Franco
432, 488 Claus, Thorsten
520 Costigan, Catherine
520 Coulter, John
527 Cremonesi, Paolo
453 De Angelis, Costantino
628 De Luca, Marco
381 Denkova, Rositsa
381 Denkova, Zapryana
59 Dhanapal, Jennifer
279, 313 Dostal, Petr
320, 327
25, 354 Dusek, Frantisek
333 Eberard, Damien
698 Eferina, Ekaterina G.
80 Eltawil, Amr B.
227 Ershov, Egor. I.
192 Felfoeldi-Szuecs, Nora
663 Fokicheva, Elena
474 Fonini, Julio
439 Fransoo, J.C. (Jan)
692 Gaidamaka, Yu.
474 Garcia, Vinicius J.
527 Garzotto, Franca
340 Gazdos, Frantisek
453 Gili, Valerio F.
596 Gonzalez-Velez, Horacio
381 Goranov, Bogdan
527, 621 Gribaudo, Marco
439 Grievink, Johan
658 Grusho, Alexander A.
658 Grusho, Nick A.
596 Grzonka, Daniel
698 Gudkova, Irina A.
11 Guenther, Willibald A.
474 Guimaraes, lochane
534, 564 Hameed, Ibrahim A.
204 Hannappel, Marc
446 He, Fang
432, 488 Herrmann, Frank
554 Hjorungdal, Rolf-Magnus
300 Holis, Radek
25, 354 Honc, Daniel
245 Hrabec, Dusan
7 Husslein, Thomas
527, 621 Iacono, Mauro
467 Ivanov, Dmitry
159 Ivlev, Anton
596 Jakobik, Agnieszka
508 Jaturanonda, Chorkaew
166 Juhasz, Jacint
172, 217 Juhasz, Peter
53 Kaczorek, Tadeusz
467 Kalinin, Vladimir

368, 460 Kang, Bong Gu
 227 Karpenko, Simon M.
 515 Khompatraporn, Chareonchai
 368, 460 Kim, Byeong Soo
 368, 460 Kim, Tag Gon
 603 Kirichek, Ruslan
 663 Kiseleva, Ksenia
 495 Kitowski, Jacek
 425 Knobloch, Roman
 596, 641 Kolodziej, Joanna
 237, 258 Kominkova O., Zuzana
 93 Kon, Cynthia M. L.
 651 Konovalov, Mikhail
 663, 669 Korolev, Victor
 676
 685, 698 Korolkova, Anna V.
 663, 676 Korotysheva, Anna
 381 Kostov, Georgi
 692 Kovalchukov, R.
 66 Krauklis, Kaspars
 287, 307 Krhovjak, Adam
 327
 495 Krol, Dariusz
 279, 347 Kubalcik, Marek
 481 Kuehn, Mathias
 685, 698 Kulyabov, Dmitry S.
 25 Kumar T., Gireesh
 231 Kuncar, Ales
 93 Labadin, Jane
 66 Lace, Inta
 698 Laneev, Eugeny B.
 413 Laue, Ralf
 508 Leerojanaprapa, Kanogkan
 33 Legato, Pasquale
 453 Leo, Giuseppe
 614 Levis, Alexander H.
 87 Liang, Chen
 453 Locatelli, Andrea
 340 Macek, Daniel
 159 Makhabat, Abilova
 33 Malizia, Lidia
 293 Maly, Martin
 621 Manini, Daniele
 579 Marini, Alessandro
 607 Marks, Michal
 333 Marquis-Favre, Wilfrid
 107 Martins, M. Rosario
 5 Matta, Andrea
 33 Mazza, Rina Mary
 502 Mellouli, Taieb
 388 Melnik, Roderick
 710 Meykhanadzhyan, Lusine
 381 Minkova, Svetlana
 425 Mlynek, Jaroslav
 692 Moltchanov, D.
 413 Mueller, Christian
 265 Musa, Harald
 152 Nagy, Laszlo
 166 Nagy, Julia T.
 166 Nagy, Balint Z.
 145 Neck, Reinhard
 467 Nemykin, Sergey
 393 Neumann, Petr
 107 Neves, Jose
 607 Niewiadomska-Szynkiewicz,
 Ewa
 18 Nigro, Libero
 227 Nikolaev, Dmitry P.
 381 Ninova-Nikolova, Nadya
 265 Nolle, Lars
 270 Notermans, Guido
 293 Novak, Jakub
 628 Oreggia, Marco
 152, 179 Ormos, Mihaly
 185
 534, 544 Osen, Ottar L.
 554, 564
 211 Patlins, Pavel
 534, 564 Pedersen, Birger S.
 114 Peirano, Lorenzo
 374 Pekar, Libor
 138 Perez, Pascal
 381 Petelkov, Ivan
 41 Petermann, Arne
 361 Petre, Emil
 333 Pham, Minh Tu
 138 Phuong, Van Hoang
 527 Piazzolla, Pietro
 495 Pietrzyk, Maciej
 107 Piteira, Ana
 237, 245 Pluhacek, Michal
 252, 258
 211 Pocs, Remigijs
 393, 399 Pospisilik, Martin

227 Postnikov, Vassili V.
 388 Prabhakar, Sanjay
 419 Privalov, Alexander
 374 Prokop, Roman
 231 Prokopova, Zdenka
 446 Qu, Rong
 587 Raho, Daniel
 495 Rauch, Lukasz
 439 Ravi, Narayen K.
 651, 676 Razumchik, Rostislav
 710
 587 Rigo, Alvis
 138 Ritter, Allison
 453 Rocco, Davide
 502 Roemer, Michael
 481 Rose, Oliver
 287, 307 Rusar, Lukas
 327
 544, 554 Rutle, Adrian
 692 Samuylov, Andrey
 544, 554 Sanfilippo, Filippo
 663, 676 Satin, Yacov
 534, 564 Schaathun, Hans G.
 520 Scheiber, Ernest
 270 Schuett, Jennifer
 18 Sciammarella, Paolo F.
 59 Sekar, P. A. J.
 432, 488 Selmair, Maximilian
 237, 252 Senkerik, Roman
 258
 685, 698 Sevastianov, Leonid A.
 138 Shanahan, Marian
 25, 354 Sharma K., Rahul
 663, 676 Shilova, Galina
 381 Shopska, Vesela
 676, 692 Shorgin, Sergey
 159 Shubat, Oksana
 138 Shukla, Nagesh
 231 Silhavy, Radek
 41 Simon, Alexander
 59 Sivaramakrishna, S. T.
 495 Slota, Renata
 467 Sokolov, Boris
 508, 515 Somboonwiwat, Tuanjai
 368 Song, Hae S.
 66 Spalvins, Aivars
 73 Spezzano, Giandomenico
 587 Spyridakis, Alexander
 425 Srb, Radek
 361 Stinga, Florin
 635 Suchacka, Grazyna
 217 Szaz, Janos
 520 Tabirca, Sabin
 628 Tacchella, Armando
 279, 287 Talas, Stanislav
 307, 327
 87 Tang, Guolei
 432 Teich, Enrico
 227 Terekhin, Arseniy P.
 121 Theerawongsathon, Prasit
 265 Thormaehlen, Holger
 658 Timonina, Elena E.
 179, 185 Timotity, Dusan
 432, 488 Trost, Marco
 419 Tsarev, Alexander
 705 Tsarev, Alexey
 231 Urbanek, Tomas
 381 Urshev, Zoltan
 439 Van Sint Annaland, Martin
 172, 217 Varadi, Kata
 100 Vasileva, Svetlana
 685 Velieva, Tatyana R.
 544 Verplaetse, Tom
 107 Vicente, Henrique
 217 Vidovics-Dancs, Agnes
 237, 245 Viktorin, Adam
 252, 258
 481 Voelker, Michael
 313, 320 Vojtesek, Jiri
 198 Walailak, Atthirawong
 121 Wasusri, Thananya
 11 Wenzler, Florian
 270 Werner, Jens
 145 Weyerstrass, Klaus
 641 Wilczynski, Andrzej
 59 Williams, Edward J.
 87 Yu, Jingjing
 481 Zahid, Taiba
 663, 669 Zeifman, Alexander
 676
 252 Zelinka, Ivan
 481 Zhou, Zhugen
 439 Zondervan, Edwin

